LEARNING AND RESTRUCTURING CAUSAL CONCEPTS

BY

ERIC GREGORY TAYLOR

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2010

Champaign, Illinois

Doctoral Committee:

Professor Brian H. Ross, Chair
Professor Gary S. Dell
Professor John E. Hummel
Professor Jose P. Mestre
Professor Daniel J. Simons

ABSTRACT

Typical studies of concept learning in adults address the learning of novel concepts, but much of learning involves the updating and restructuring of familiar concepts. Research on conceptual change explores this issue directly but differs greatly from the formal approach of the adult learning studies. This paper bridges these two areas to advance our knowledge of the mechanisms underlying concept restructuring. The main idea behind this approach is that concepts are built on causal-explanatory knowledge, and hence, models of causal induction may help to clarify the mechanisms of the restructuring process. A new paradigm is presented to study the learning *and revising* of causal networks. Experiments 1 and 2 showed that learners' prior beliefs about the causal relations in a domain affected their hypotheses as they began to infer the correct causes. First, when the prior learning suggested evidence against some of the incorrect causes, this helped learners to focus on the correct causes later in learning. Second, the prior causal beliefs were difficult to give up, and they biased learners away from the correct causes that competed to explain the same effects. Experiment 3 showed that learning by intervention, as opposed to observation, affected the concept restructuring process in different ways, depending on what interventions were chosen and by whom. People choosing their own interventions revealed a confirmation bias to preserve their prior beliefs, but those with no choice used the same interventions to disconfirm their prior beliefs. Taken together, these studies represent the beginnings of a larger research effort to use the analytic tools from causal induction to reveal the mechanisms behind larger shifts in knowledge, as evidenced by developing children and experts.

ACKNOWLEDGMENTS

Foremost, I acknowledge my parents and family for their love and support. I cannot imagine being here—emotionally, intellectually, and of course quite literally—without them. I also thank my family at UIUC in the Department of Psychology for sharing their many inspiring qualities—curiosity, warmth, humor, patience, and mutual interest in great food and drinks.

I am profoundly grateful to Brian Ross—for his wisdom and guidance, and for offering without any hesitation countless hours of his work and personal time to help me succeed. Thanks as well to Art Markman for showing me how much fun this job is and for providing many opportunities, both in his lab and at UIUC. My two "part-time" advisors, John Hummel and Aaron Benjamin, were also an important part of my graduate training. John, you forced me to be concrete (and to scream when I make a discovery!), and that made me a better scientist. Aaron, you gave me a well-rounded education, complete with research ideas, music recommendations, golf tips, and advice for how to be happy at work and at home.

Several of my friends from Texas and Illinois deserve special recognition for spending so much time with me (with minimal coercion): Stefania, Chris, Erik, Micah, Mike Bartnett, Nick, Audrey, Derek, Eamon, Erica, Gary, Mike Diaz, Rachel, and Steve & Erin. Special thanks to Stefania, whose affection and amazing Italian dinners made the past two years very special.

Finally, many others gave helpful comments on the specific research below, including Jose Mestre, Woo-kyoung Ahn, Bill Brewer, Gary Dell, Noah Goodman, Frank Keil, Tania Lombrozo, Bob Rehder, Pat Shafto, Dan Simons, the Ross and Hummel lab groups, Dan Navarro, and three reviewers from the Cognitive Science Conference.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

Concept learning is an incremental process. We learn each new concept for the first time only once, and often our initial understanding is flawed. The remainder of learning involves the updating and revising of previous knowledge. The critical implication—that most learning is actually the refinement of familiar concepts—runs counter to the traditional approach in studies of concept learning, which has focused on the learning of entirely novel concepts. Many open questions remain on the nature of conceptual restructuring.

The goal of this thesis is to better understand the mechanisms of concept restructuring by forging a connection between the traditional work on concept learning and the literature on conceptual change. Although these two areas differ greatly (in everything from goals to dependent measures), this proposal builds on recent work that highlights their commonalities.

Studies of conceptual change outline the process of knowledge restructuring in broad strokes, using different age groups and expert populations to infer the critical transitions. The different types of change are inferred from the structural differences in theories at different stages (e.g., Carey, 1985; Chi, Slotta, & De Leeuw, 1994), or by revealing students' biases in evaluating anomalous data (e.g., Chinn & Brewer, 1993; Koslowski, 1996; D. Kuhn, Amsel, O'Loughlin, & Beilin, 1988). To support these claims, authors have often focused on specific real world domains like biology and physics.

The work on conceptual change differs dramatically from the traditional studies of concept learning in adults, despite much overlap in interests. First, the adult work has focused on artificial concepts and domain-general laboratory paradigms, explicitly to avoid the role of prior knowledge. Second, the questions commonly asked are narrower in scope, motivated by formal

models of the representations (prototype, exemplars, rules) and processes underlying basic tasks like classification, inference, and category-based induction (Murphy, 2002).

A complete understanding of concept learning and restructuring requires explanations from both levels of analysis. I suggest that recent work developing the theory view of concept representation (Murphy & Medin, 1985; Wellman & S. A. Gelman, 1992) serves as a linkage between these levels. The theory view states that concepts are built upon networks of causal-explanatory knowledge. This knowledge affects performance in laboratory-based learning tasks (Murphy, 2002) and plays a role in the learning and development of real world concepts where conceptual change effects are typically shown (Vosniadou, 2008). Assuming that concept learning amounts, in large part, to the learning of causal relations, then models of causal reasoning provide the requisite theoretical tools for understanding the basic mechanisms of concept learning (Griffiths & Tenenbaum, 2009; N. S. Kim & Ahn, 2002a; Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2008; Rehder, 2003a, 2003b) and potentially also conceptual change.

Few previous studies address this linkage directly, however. The goal of the proposed work is to bridge the formal approach to concept learning with the conceptual change literature by showing that some aspects of conceptual change might result from basic causal learning mechanisms. In Experiments 1 and 2, I develop and test a new paradigm to study concept restructuring, extending the basic causal structure learning task (Lagnado & Sloman, 2004, 2006; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003) to cases where learners not only acquire a novel concept but are also prompted to restructure that concept later in learning. Given that we currently know much about the learning of novel concepts defined by causal relations, the

current studies will address how one's extant causal knowledge affects continued learning and restructuring.

In Experiment 3, I move forward with this paradigm by exploring the role of other helpful tasks that may promote conceptual change. I focus on at active learning task called "intervention," where learners manipulate the status of one property of the causal system and then observe the effects on other properties. Previous work shows that interventions improve learning by helping to disambiguate alternative causes (Steyvers, et al., 2003), precisely the task of learners needing to restructure their concepts. Interventions also provide a direct test of causal understanding, separating association-based learning from true causal inference. Experiment 3 builds on previous work by testing the role of intervention in revising one's incorrect prior beliefs. To foreshadow, different types of intervention are more effective than others, and some are actually detrimental. I discuss implications for pedagogy and classroom learning techniques.

In sum, this work extends research in the growing area of causal induction and bridges the formal approach to concepts with studies of conceptual change toward an integrated view of concept learning and restructuring. Taken together, these experiments suggest many opportunities for future work to engage this bridge and find techniques that improve learning and conceptual change.

CHAPTER 2: RESTRUCTURING OF REAL WORLD CONCEPTS

The primary task for conceptual change research is to map out the differences in people's concepts over time—to determine what the concepts are changing to and from. This map can then be used to develop and test theories as to how people transition from one state of conceptual knowledge to the next. A common technique is to interview different people either at different ages or with varying degrees of expertise in the topic of interest, and infer from their responses the differences in latent conceptual knowledge. This approach has typically addressed people's concepts for real world domains, like children's understanding of biology and students' knowledge of science topics. The results have been highly revealing about the longitudinal nature of concept restructuring—what types of transitions occur, how sudden or gradual they are, and whether learning in one domain is independent from another. The goal of this section is to review two areas, cognitive development and science learning, to show what concept restructuring looks like, and concurrently, to argue that changes in causal knowledge are at the core of conceptual change in both areas.

2.1 Cognitive Development

Piaget laid the foundation for work on cognitive development. He argued that cognition develops in stages, where each stage is defined by progressively more advanced domain-general abilities (Piaget, 1952). For example, children were presumed to reason without systematic cause-effect schemata until roughly first grade. Stage theory incited more than a half-century of rigorous debate over the nature of development (see Flavell, 1996), and although Piaget's ideas were revolutionary, they are now effectively antiquated by work showing that children do possess coherent systems of beliefs from an early age (Gopnik, 1996). Recent views are organized into three domains for which children demonstrate competence in reasoning even at an

early age: naïve physics, psychology, and biology (Wellman & S. A. Gelman, 1992). Although each domain could be reviewed extensively, I'll focus on naïve biology in particular, since this is where many theoretical claims relevant to the nature of conceptual change are clarified. Naïve physics will also be addressed in the following section on science learning.

2.1.1 Naïve biology

Much debate over the development biological concepts concerns the role of causal-explanatory knowledge. Two important debates have addressed (a) whether children have separate domains of explanatory knowledge (as opposed to general world knowledge used across domains) and (b) to what extent the fundamental nature of the explanations change, possibly independent of domain.

2.1.1.1 Changes in domain knowledge

A recent landmark in the post-Piagetian tradition is Carey's work on naïve biology, which attributes (controversially) knowledge of just two domains to young children—physics and psychology. Carey (1985) argues that children under about 10 years old lack a basic understanding of biology, and as a consequence, tend to perceive biological events by reference to humans and their psychological motives. For example, if asked to explain the role of quintessentially biological events, like eating or sleeping, a young child will cite hunger and sleepiness as causes, seldom mentioning the body's need for nourishment. Biological concepts eventually develop through bootstrapping from the psychological domain (Carey, 1999).

Carey's (1985) studies of property extensions with young children offer compelling evidence for this view. Children were told that some entity, typically an animal or plant, possessed a novel property (e.g., a "golgi") and then asked to say whether they thought a new

entity would also have that property. The inferences made suggest how a child's knowledge of biological kinds is organized. In one study, 6-year olds were told that dogs and flowers have a "golgi" and asked what other things in the world may also have a golgi. Adults rarely extended the property golgi to non-living things—noting the fact that if animals and plants share it then it must be common to living things—but 6-year olds commonly did so. Based on this finding, Carey suggests that what children see in common among animals and plants is not their fundamental biological nature, but instead, that they are merely both kinds of objects.

Other studies from Carey (1985) showed that children supplant their lack of biological knowledge by using humans as their default referent category. For example, 4-year olds said that the organs of humans (e.g., a spleen) are probably also inside of bees, but strangely, that organs of stinkbugs are less likely inside of bees. In contrast, adults and older children showed reasoning based on the inclusion of species in common taxonomic categories (e.g., mammals, insects). It seems that young children lack these basic taxonomic distinctions, so they refer to humans when possible and are simply unsure otherwise.

As a side note, one might infer from this bias that children's preferred, or default, style of reasoning is explanatory, not taxonomic. Plenty of evidence suggests that even infants have rich intuitive abilities to explain behavior in terms of psychological motives (Woodward, 2005). Children's early tendency to extend properties relative to category human is perhaps a misguided attempt to apply their expertise in causal/explanatory reasoning to a normatively taxonomic task.

Carey's view is not without critics, but even those who do attribute rudimentary biological concepts to children show that it's structured very differently from that of adults. Inagaki and Hatano (1993) found that 6-year olds do make some distinctions between biology and psychology, since they preferred explanations for eating and breathing based on the

workings of inner organs (i.e., biological entities) to explanations based on psychological motives. However, the same 6-year olds did not yet prefer mechanistic explanations (e.g., the lungs take in oxygen and change it into carbon dioxide) to explanations centered on the needs and intentions of the organs (e.g., the chest takes in vital power from the air). This contrasts with 8-year olds and adults who did prefer the mechanistic explanations.

Inagaki and Hatano suggest a weaker version of Carey's view—that children do have knowledge of biological kinds but view them as primarily achieving the goals of an agent; for young children the agent is simply the biological organ. Note, however, they have not weakened the view that causal models of behavior drive many of children's early attributions. What changes with development is the nature of the agents that fill the causal roles, but not the role of causal reasoning in general.

2.1.1.2 Changes in the nature of explanation

Debate continues over children's early conceptions of biology, and consequently, some have taken a different approach less centered on the content of children's beliefs and more on the types of explanation deemed relevant. A fact can typically be explained in a number of ways, but young children often rely on "teleological" (i.e., functional) explanations in particular (for a review, see (Lombrozo & Carey, 2006). For example, when asked why plants are green, second graders prefer the teleological answer, "because it is better for the plants to be green and it helps there be more plants" to the mechanistic one, "because there are little tiny parts in plants that when mixed together give them a green color" (Keil, 1992).

Relating back to the debate over children's naïve biology, Atran (1995) and Keil (1992) argued that this preference was specific to children's reasoning for biological kinds and living things. However, later evidence showed that children also give teleological explanation of

artifacts (Kelemen, 1999). Kelemen's view is more general, that children are "promiscuous" teleological explainers, and this stems from their goal-directed, agent-based understanding of the world. Children explain not only artifacts but also patently non-functional objects in terms of their role in goal-directed events (e.g., animals are "for going to the zoo" and plants are "for watering the plants.") Of course, later in life adults sensibly restrict their teleological explanations to artifacts and sometimes to goal-relevant parts of biological kinds (e.g., eyes, opposable thumbs).

Lombrozo and Carey (2006) argued that as children develop what might be changing is their relative weighting of an item's function compared to other non-functional background knowledge. Based on the knowledge one has (and not necessarily their age or the domain), either functions or mechanisms may seem more important and bias the explanations accordingly. Thus, if a young child's understanding of an event was (atypically) non-intentional, they would explain the event with reference to the mechanisms that instantiated it rather than the post-hoc function it may appear to serve.

To test the idea that knowledge, not domain biases, affects the types of explanation given, Lombrozo and Carey (2006) had adults explain an ambiguous scenario where a tourist cave was enlarged, either on purpose to attract more tourists or by accident when reinforcing the caves for safety. They found that even when participants were told that the cave's size has a function (e.g., tourists prefer large caves), participants whose background knowledge suggested that the cave enlargement was an accident were more likely to endorse a mechanistic explanation for the size. Interestingly, the same effects were obtained for explanations of artifacts (e.g., satellite dishes) and natural kinds (e.g., caves), casting further doubt on domain specific preference for teleological explanation.

These findings suggest that the contents of one's causal knowledge can affect the structure of explanation, in so far as it affects one's view as to how and why the causal system originated in the first place. Returning to naïve biology, children will learn some day that the lungs help to oxygenate blood, thus realizing that the lungs are not really an agent needing "vital powers" but instead, one of many implements used to support biological functions. Combined with Carey's work, the studies of explanation shows that children's developing causal knowledge is a crucial factor driving the major developmental shifts—by affecting the apparent organization of their biological concepts and the types of explanations that are preferred.

2.1.2 Causal inference

Much recent work continues to emphasize the role of causal reasoning and explanation in development and has asked an even bolder set of questions about the competency of very young children to think about causes and effects. Causal reasoning is notoriously difficult, and not merely as a practical matter, but also from a theoretical perspective when trying to understand how and when, in principle, one can infer causation from observations (Hume, 1978). Adults reason in accord with some formal principles (Cheng, 1997; Griffiths & Tenenbaum, 2005), but struggle when learning complex causal structures (Steyvers et al., 2003) and when making counterfactual inferences (Meder, Hagmayer, & Waldmann, 2009). Hence, recent findings that even young children appear to make causal inferences based on covariation statistics are all the more surprising.

In one study, Gopnik and colleagues (Gopnik, Sobel, Schulz, & Glymour, 2001) showed that 3- to 4-year old children could selectively intervene on a causal system to bring about an effect. The ability to intervene separates true causal inference from association-based learning, since the behavior initiated by a causal intervention requires no training or observational

precedence. In one task, children watched as an experimenter placed one of two boxes onto a platform, one of which—the "blicket"—would cause music to play. On the final trial, both boxes were placed on the platform and children were asked to "make the music stop." Although the children had never operated the blicket machine, nor had they seen the experimenter perform an action specifically to make the music stop, 75% of children selectively removed the blicket and not the other box.

(Gopnik et al., 2004) account for these findings and others with their view that children learn and reason with "causal maps," or abstract coherent representations of cause-effect relationships. Specific causal maps are learned by observation and interactions with real world causal systems, but their governing principles are domain-general, corresponding roughly to the assumptions underlying causal Bayesian networks (Pearl, 2000). They emphasized that causal maps are less general than connectionist approaches to development (Elman, Bates, Johnson, & Karmiloff-Smith 1996; Doumas, Hummel, & Sandhofer, 2008), since they apply exclusively to reasoning about causal systems, but are more general than nativist approaches, which assume separate core systems for different domains; e.g., the number system (Leslie, Gallistel, & R. Gelman, 2008), the object-mechanics system (Spelke & Kinzler, 2007).

The specific claims of the Gopnik et al. (2004) view are still being explored and tested, but the general point finds agreement with others in the field—that as children grow older, they represent ever-more complex systems of causal relations between entities and events in the world (Baillargeon & Goswami, 2002; Cohen & Oakes, 1993; Leslie & Keeble, 1987; Wellman & S. A. Gelman, 1992). Without reviewing the developmental work in more depth (though nearly all of it is relevant to conceptual change), I note that concept restructuring centrally involve changes in the structure and usage of causal knowledge. First, children supplant their lack of taxonomic

knowledge with their causal understanding of human's psychological motives and observed similarities between humans and other species. Second, the same proficiency in understanding humans and their goals leads children to overuse teleological explanations. This preference shifts when further knowledge is accrued that suggests the role of causal mechanisms, rather than functions. Third, recent work shows that young children's understanding of events is truly causal, since they can reason about causal interventions that were directly associated with the desired outcome.

The next section builds on the idea that changes in causal knowledge are at the core of concept restructuring. I review some of the classic work and current debates in the area of science learning. These studies are continuous with the developmental work, but differ in their extensions to more complex domains (e.g., physics, mathematics), adult learners as well as children, and how knowledge changes as a function of expertise.

2.2. Science Learning

Research on science learning grew from the classic studies of expertise, which focused mainly on non-scientific tasks. For example, in a seminal paper, Chase and Simon (1973) showed that expert chess players perceived larger "chunks" of the chessboard, implying that expertise affects how one represents a given domain, and consequently, how they are able to interact with it and problem solve. Chess is a difficult and fascinating game, but it pales in complexity to scientific topics like mechanics or molecular biology. Expertise in science is fundamentally different, especially regarding causal knowledge—the goal of a scientist is to understand the relations between natural entities and their environment, which in practice, involves more than a collection of strategies to navigate a finite problem space. Scientists first learn to solve problems, but then learn to ask question, create new problems, and develop novel

techniques to obtain the solutions. The studies reviewed below show how students develop these skills, beginning with problem solving, and then shifting to problem representation and understanding. I emphasize that much of what develops on the path to expertise is systems of causal knowledge.

2.2.1 Categorization and problem solving

A useful technique for solving problems is to first categorize the problem, and then apply the knowledge of the category to obtain the solution. Students commonly use this approach, especially given that textbooks chapters are often organized by problem categories. For example, problems on "integration" may be grouped into one or two chapters, and on the final exam, knowing the category of "integration problems" brings to mind the rules of calculus needed to find the solution. Many previous studies have examined how problem categories develop with expertise, and the results offer important insights into the process of conceptual change.

A famous study by Chi, Feltovich, and Glaser (1981) showed that for physics problems, novices categorize based on the problem's surface features (e.g., problems with an "inclined plane"), whereas experts categorize based on the underlying theoretical principles (e.g., problem that instantiate Newton's second law). In some sense, these findings are similar to those of the chess studies—novice physicists simply do not perceive the more complex relations between problem features.

Although novice categories in physics are objectively wrong, in practice, they may be somewhat right but for the wrong reasons. Mayer (1981) observed that textbook problems related by a common theory often involve a set of "usual suspect" objects, meaning that problem categories based on these objects will predict the correct solution path more often than chance. Interestingly, even experts seem to take advantage of these object correlations; Blessing and

Ross (1996) showed that experts were more accurate on problems with the usual objects. These findings suggest that a transition may take place where learners begin to see the underlying principles, but when in doubt, revert to the surface features.

Other research speaks to how this transition might occur. One idea is that novices use the surface features to make comparisons between problems within a category. For problems that share surface features and the deeper principles, the comparisons help to identify the principles and incorporate them into the domain knowledge (Ross, 1984; Ross & Kennedy, 1990).

Another idea suggested by Carey (1999) is that students may initially understand the deeper principles, but do not know how apply them. However, with experience, the principles grow in usage as they help resolve contradictions posed by feature-based problem solving; e.g., for problems with multiple features that suggest different categorizations.

Both ideas suggest that expertise for complex domains develops in steps, from knowledge based on simple features to deeper principles. Experiments 1 and 2 below address this idea with relations to causal learning. I create a stepwise learning approach by having people first note a correlation and then use it as a constraint when building more a complete causal understanding. The next section shows how this type of transition in causal knowledge relates to studies of concept development in students and scientists.

2.2.2 Naïve theories vs. knowledge in pieces

The study of problem categories is just one approach to understanding the learning of science concepts. Other research threads have examined the structure of scientific knowledge by considering the representations in more depth and evaluating these with student testimonies taken from classroom settings. Current debate revolves around two competing views of concept representation: the "naïve theories" view and the "knowledge in pieces" view. The goal of this

section is to discuss the evidence for both views, to point out their distinctions, and to emphasize the role of causal knowledge in each. Interestingly, I show that when the views are recast purely in terms of causal representations, their differences are minimized, suggesting a major benefit to viewing conceptual change as developments in causal knowledge.

2.2.2.1 Naïve theories

The naïve theories view was founded on the analogy between conceptual change in individuals and in the history of science. This analogy holds that prior to instruction students possess a coherent set of explanatory concepts for particular domains (much like in Carey's view) and that these bear a resemblance to early scientific theories. For example, naïve concepts of force are often similar to the outdated "impetus theory," in which objects were believed to inherit force when set into motion (McCloskey, 1983). The analogy also implies that students develop into experts much as scientific theories mature over time—the initial theories are strongly held by both students and scientists, and when change does occur the later concepts are often not translatable in terms of the prior concepts (Carey, 1999; T. S. Kuhn, 1970).

Much evidence for the naïve theories view comes from interview studies with children and students. In one study, Vosniadou and Brewer (1992) examined elementary school children's concepts of the earth. They found that most of the children interviewed (ages 6-11) possessed one of several distorted, but coherent, mental models of the earth (akin to diagrams). One model depicted earth as a sphere but with the top hemisphere hollowed out to allow for the fact that people walk on a flat surface. The naïve theories view permits the possibility that different children may have different models, but says that all of a particular child's responses should be consistent with their personal model. Indeed, nearly 80% of children responded in a consistent

way. A later study by Ionnides and Vosniadou (2002) replicated this finding in the domain of physics.

Viewing students as naïve theorists leads to ideas for how to foster conceptual change. Specifically, instructors should confront students, point out their theoretical misconceptions (much as scientists would identify flaws in published papers), and suggest methods for replacing old theories with the scientific ones. Certain techniques do help to remove the misconceptions in one's prior theories (e.g., Clement, 1993; Thornton & Sokoloff, 1990), but in general, revising a theory is difficult for both students and scientists (Carey, 1988; T. S. Kuhn, 1970), and there may be downsides to telling students that their prior conceptions are entirely wrong. For one, students could develop the view that physics, and science more generally, is usually counter-intuitive (Smith III, diSessa, & Roschelle, 1994).

2.2.2.2 Knowledge in pieces

The knowledge in pieces (KIP) view suggests an alternative theory and alternative recommendations for teachers of science. The central claim is that students do not have organized, coherent theories, but rather the opposite. Early scientific concepts consist of numerous disorganized pieces of intuition, accumulated through experience in the physical world and applied when making sense of events and observations. diSessa (1993) called these intuitions "p-prims" (for phenomenological primitives). An example p-prim is "more effort begets more result," which is phenomenological in the sense that we commonly observe events in the world that confirm it and is primitive in the sense that it seems to require no further explanation; it is evidentially true. The activation and usage of p-prims also differs from naïve theories. Unlike naïve theories, p-prims are not retrieved systematically for events in the same domain. The p-prim that gets activated for a given observation is based on contextual variables, including many

factors not inherent to the domain itself (cf. disessa, Gillespie, & Esterly, 2004; Thaden-Koch, Dufresne, & Mestre, 2006).

Children and novices err in their understanding of physics by overextending p-prims to cases where they don't apply. As students develop, the relations between p-prims and events are revised so that given a problem of a specific type, a particular p-prim is likely to be activated. This points to another major difference between the KIP and naïve theory views. According to KIP, conceptual change involves the refinement of p-prim activation but not the elimination of p-prims (as the naïve theory view requires replacement of bad theories). Learning involves simply developing and refining the right intuitions (Smith III et al., 1994).

Distinguishing between these two views has been difficult, since the classroom-based evidence is often indirect and the commonly used data-coding procedures are susceptible to the biases of the researchers. In fact, the same data were published twice with analyses from each viewpoint leading to opposite conclusions (disessa et al., 2004; Ioannides & Vosniadou, 2002). A central goal of this thesis is to emphasize the common role of causal knowledge in differing views on concept development and science learning. The next section provides a re-description of the science learning views in terms of causal models, implying much like the Gopnik et al. (2004) theory did in the context of child development, that conceptual change amounts in large part to the learning and restructuring of causal relations.

2.2.3 Science learning as causal structure learning

Both the naïve theories view and the KIP view suggest that causal knowledge plays a special role in the learning of science (diSessa & Sherin, 1998; Ioannides & Vosniadou, 2002). Causal networks are part and parcel of the naïve theories view, since theories are defined literally as sets of causal explanations (Wellman & Gelman, 1992). Given that young children already

seem to have intuitive theories organizing their knowledge of science topics, and that these are typically quite off the mark, the development of scientific knowledge must involve theory change. To see how conceptual change might occur according to the naïve theories view, I'll distinguish two types of theory change using Carey's (1988) notions of weak and strong restructuring. I discuss the role of causal learning in each, showing that both types of conceptual change over theories rely centrally on changes in causal knowledge.

2.2.3.1 Types of theory change

According to the naïve theories view, human concepts develop much like scientific theories over the course of history. Carey (1988) unpacked this analogy to test whether children's concepts actually develop like scientific theories. She designed a set of criteria to distinguish "weak" and "strong" restructuring, where only the latter describes concept development in the history-of-science sense.

First, weak restructuring is not like theory change in the history of science. It involves the mere updating and reorganizing of existing causal networks by adding new concepts and new causal relations. For example, novice physics learners might represent the conditional, "no motion without force," but experts would say, "no *acceleration* without force." This requires the addition of the new concept "acceleration" plus a new causal link from force to acceleration, nothing further.

Strong restructuring adds three new qualifications: (1) changes at the level of individual concepts, (2) changes in the nature of the explanations used, and (3) changes in the domain to which these new concepts and explanations apply. Carey (1988) demonstrated how to test for these by citing evidence for strong restructuring in children. First, in older children the emergent concept of "living things" is not merely a disjunction of animals and plants, since these

categories are now seen as fundamentally alike in some new way (having life). Thus, (1) is satisfied because change occurs for the individual concepts "animal" and "plant." Second, requirement (2) is met because much work shows that children shift in the types of explanations they give for biological processes, like eating and sleeping, only later citing biological mechanisms, like the need for such processes to maintain life. Requirement (3) is met because the theories of older children's now explain a new domain—that of biology. Carey (1988) argues that earlier theories do not explain biology, per se, because the behaviors they account for do not pick out essentially biological tasks like providing nourishment; instead they focus on social aspects (e.g., for eating, the differences in types of meals and when it's okay to eat with your fingers). Thus, having met these requirements 1-3, Carey argues that children do undergo strong theoretical change.

2.2.3.2 Theory change is change in causal knowledge

In light of these finer distinctions in types of theory change, the goal of this section is to recast both types of restructuring (weak and strong) in terms of the learning and revision of causal models. I being with weak restructuring, the simpler form, which is relatively easy to recast. According to Carey, weak restructuring is literally the addition of new concepts (i.e., learning about new types of properties or events) and their integration with the existing concepts through added causal relations; hence, no work is needed on my part to recast the view in causal terms. This example of force and acceleration demonstrates this shift, as mentioned earlier. One simply forms the new concept of "acceleration" events (as distinguished from motion) and then posits a new causal relation between acceleration and force.

Strong restructuring is more complex, involving the three qualifications from above, and requires further assumptions to cast purely in terms of causal knowledge. I address the

qualifications one at a time. First, strong restructuring implies change at the level of individual concepts. To address this point, and to ensure generality, I adopt a view that concepts are defined both *externally* and *internally* (Goldstone, 1996; Goldstone & Rogosky, 2002). External structure refers to the relations that concepts participate in with other concepts, so that for example "force" is defined partly as "that which acceleration cannot exist without." These external relations can be learned via the processes of weak restructuring as shown above—by adding concepts and connecting them to existing causal structures.

Internal structure refers to the causal mechanisms that give rise to instances of the concept (Kim & Ahn, 2002; Rehder, 2003a, 2003b). For example, living things are instantiated by a causal scheme involving DNA, innate biological parts and behaviors, and adaptive methods for survival. To represent these internal relations, developing children must do two things. First, they need to learn the new sub-concepts entailed, such as "DNA" and "survival." Second, they need to connect these concepts with causal relations to match the generative mechanisms that instantiate members of the class "living things." As can be seen, this combination of steps is again equivalent to the process of weak restructuring. Hence, change at the level of individual concepts occurs by "weak" restructuring of concepts' external and internal structures.

Strong restructuring has two further qualifications—changes in the nature of explanation and changes in the explanatory domain. I argue that these may also be done via the processes of weak structuring. First, change in the nature of explanations, e.g., from psychological to biological, may arise partly with the addition of the new concepts, like "nourishment" and "survival," which may be taught, and by explicating their causal relation to known biological events like eating and sleeping. Once established, these new relations may gain in preference as other developmental changes occur that shift the child's view of what types of explanations are

relevant (Keil & Newman, 2008). Thus, the nature of explanation shifts by adding knowledge and accruing evidence in favor of the new forms of explanation.

Second, with the addition of concepts like as nourishment and survival, plus their linkage to known behaviors, qualification (3) comes for free—children's theories are now extended to new events in the domain of biology, including not only the new biological sense of eating and sleeping, but also other events that nourishment and survival explain. For example, children could now understand that some people sunbathe to obtain vitamin D, and that competition for survival among species accounts for natural selection.

Hence, we may cast both weak and strong restructuring in terms of causal models, with no added assumptions for weak restructuring, and by assuming that changes in the nature of individual concepts occur either externally—by adding new concepts and integrating them with current ones through new causal relations, or internally—by editing the causal mechanisms that give rise to conceptual tokens. Strong restructuring is reducible to weak restructuring, because requirements 1-3 are met with addition of new concepts and new links to prior knowledge.

2.2.3.3 Changes in p-prims are changes in causal knowledge

In spite of the alleged differences between KIP and the naïve theories view, causal relations play a central role in both. On the KIP end, diSessa and Sherin (1998) incorporate causal networks into their notion of "coordination class," which they say is a type of concept commonly evidenced by students of science. A coordination class is defined by the combination of strategies and intuitive knowledge (p-prims) students use to understand a given class of physical events. To demonstrate, the coordination class for the concept "force" involves (a) strategies for what to attend to in a given observation when trying to determine the force

involved, and (b) a causal network consisting of the many p-prims that help one to understand force and its relation to other concepts.

According to diSessa and Sherin (1998, p. 1178), "the causal net is, in fact, the primary locus of difficulty in learning concepts of schooled physics." Other aspects of learning, such as the development of efficient "readout strategies" for perceiving physical concepts are said to evolve hand-in-hand with the causal network, as it becomes more systematically organized. But these are not the source of the problem, since people have natural abilities to perceive physical concepts (like forces and their causal outcomes; Michotte, 1963). These perceptions are simply misguided in novices due to their flawed causal knowledge.

Given these theoretical claims, the KIP view channels precisely the claim of this thesis, as did Carey's views of restructuring—that the lion's share of science learning involves the refinement of one's underlying causal knowledge. Hence, I consider the basic tenets of KIP theory in agreement with that of the theory view, in so far as conceptual change in both amounts in large part to the learning and revision of the causal structures that underlie our concepts. This view is also in accord with the work by Gopnik et al. (2004) described above in the review of work on cognitive development.

2.3 Summary of Research on the Restructuring of Real World Concepts

Reaching across the work in cognitive development and science learning, many of the empirical markers of conceptual change are evidently the result of changes in the structure and application of one's causal-explanatory knowledge. In developmental studies, children show a change in the types of explanations given for biological events. Early explanations stem from children's understanding of humans and their motives, extending the notion of intentionality to non-intentional things like bodily organs. As children learn new causal facts (e.g., that lungs

provide oxygen to blood), their explanations become more focused on causal mechanisms. Crucially, the role of causal knowledge is invariant across concept development; what changes are the bodies of causal knowledge, and consequently, the nature of the causes deemed relevant.

Other recent work in development shows that children's knowledge representations are truly causal, and that children have surprisingly advanced predictive abilities in line with the assumptions of formal models of causal learning. In controlled experiments, even young children show the ability to intervene on a causal system to bring about new results that were never previously observed. Such interventions are a trademark of causal reasoning.

Research in the area of science learning offers complementary evidence that conceptual change largely entails changes in causal knowledge. One change occurs in the organization of problem categories, which help students to solve problems in complex domains. At first, novices identify the superficial relations between problem aspects and solution paths. Although often misleading, these novice categories appear to lay the foundation for later learning of the underlying causal/theoretical principles.

Other more general views of concept representation stake claims as to how children and students develop into expert scientists. Two views in particular—the naïve theories view and the knowledge in pieces (KIP) view—are commonly treated as vehement competitors, but I argued that the apparent differences vanish when learning is cast in terms of the development of causal knowledge. This last result is especially promising for the view that conceptual change amounts in large part to learning and restructuring system of causal knowledge.

CHAPTER 3: CONCEPT REPRESENTATION AND CATEGORIZATION

In the previous sections, I discussed research in child development and science learning showing that concepts are tightly integrated with people's intuitive causal knowledge. This section reinforces those claims and extends them to the domain-general literature on concept learning in adults. Recent views highlight the role of causal knowledge and propose integrated models of causal induction and concept learning that may help to clarify the mechanisms involved in conceptual change. These mechanisms are addressed at a finer level of analysis than the work in development and science learning.

3.1 The *Other* Theory View

Let's start with a fresh example of a causally structured concept from a non-scientific domain. Consider one's concept of the political party "liberal." Liberal includes not only the properties of specific liberals (e.g., Birkenstocks, long hair, community oriented) but also the set of principles that liberals live by and causal explanations for how those principles seem to affect behavior (e.g., voting, spending)[1]. Murphy and Medin (1985) argued that causal knowledge of this sort is crucial for understanding how our concepts cohere—why some groups of items and events "go together" better than others. They offered their view, again called the "theory" view, as an alternative to the extant accounts of conceptual structure based on similarity.

Similarity-based views claim that concept representations are built upon lists of properties that tend to occur for instances of a given concept; e.g., our concept of bird includes the properties that most birds have, like wings and ability to fly (Rosch & Mervis, 1975). Models based on similarity assume that concepts are primarily used to classify new instances. To the

[1] This example also clarifies the common treatment of "concepts" in the psychological literature as the mental representations that identify the meaning of categories, such as "liberal" or "conservative."

extent that a new item is similar to concept A and less similar to other concepts B, C, etc., the item is an A (Kruschke, 1999; Minda & Smith, 2002; Nosofsky, 1986).

However, much current research shows that the notion of similarity is too unconstrained to adequately describe conceptual knowledge; what counts as "similar" shifts dramatically depending on the context (Medin, Goldstone, & Gentner, 1993), and crucially, also causal knowledge (Medin & Shoben, 1988; Rehder, 2006). For example, Medin and Shoben (1988) found that people rate grey hair as more similar to white hair than to black hair, but grey clouds as more similar to black clouds than white clouds. Presumably, these effects are driven by the causal knowledge that attributes both white and grey hair to aging, and grey and black clouds to water retainment. A similarity-based account of conceptual structure could, at the whim of the researcher, weight these features differently across different contexts, but does not provide a rationale for doing so.

3.2 Theory-based Models

Viewing concepts as knowledge-based resolves many problems with the similarity view, at least in verbal argumentation, but designing a model to formalize the view has proven difficult. Formal advancements of the theory view have appeared only recently, borrowing heavily from the causal reasoning literature. One view, the "causal status" view, holds that more typical instances of a concept are those with the "deeper" features, where depth is measured by how nested the feature is in the causal chains that represent the concept (Ahn, N. S. Kim, Lassaline, & Dennis, 2000; Sloman, Love, & Ahn, 1998). If bird DNA causes wings to develop, which causes birds to fly, then a new animal with bird DNA that has been injured and is perhaps missing a wing (and thus cannot fly) is still a "better," or more typical, bird than a human with mechanical wings and a rocket engine to enable flying.

A major virtue of this work is that causal status appears to be an important factor in people's usage of real world concepts. For example, physicians' concepts of mental disorders are represented causally, despite the fact that the Diagnostic and Statistical Manual of Mental Disorders (DSM) suggests a non-causal organization (N. S. Kim & Ahn, 2002a). Further, physicians make diagnoses according to the presence of deep features, and the same findings hold for lay people's naïve concept of mental disorders (N. S. Kim & Ahn, 2002b), suggesting a persistent influence of causal knowledge throughout learning and the development of clinical expertise.

Another view on the role of causal knowledge in concept representation is Rehder's (2003a, 2003b; Rehder & S. W. Kim, 2006) "causal model theory" (CMT). Like the causal status view, concepts are presumed to consist of networks of causal relations, but distinctly for CMT, typicality is judged more holistically and not based solely on the depth of features in the causal network. In CMT, items belong in the same category to the extent that they are likely to have been *generated* by the same causal mechanisms. This likelihood is computed with equations provided by the causal induction literature. A quick primer on the induction literature will help to clarify these equations and their accompanying theoretical claims.

3.2.1 Probabilistic causal induction

Thinking about concepts as generative causal mechanisms carries with it a number of useful models that help to formalize the underlying problem of causal induction. Traditionally, causal induction is cast as the process of inferring the strength of the cause linking two events based on the pattern of covariation between those events. An early model, "delta-p" (ΔP; Jenkins & Ward, 1965), suggests that two events are causally related to the extent that the probability of the effect given the presence of the cause is greater than the probability of the effect given the

absence of the cause: $\Delta P = p(\text{E}|\text{C}) - p(\text{E}|\sim\text{C})$. In later work, Cheng (1997) modified this expression to account for differences in people's ratings of causal strength with a common ΔP but varying $p(\text{E}|\sim\text{C})$ (Buehner, Cheng, & Clifford, 2003). Cheng's modified expression, termed "causal power," states that people's attributions of causal strength vary according to $\Delta P / [1 - p(\text{E}|\sim\text{C})]$. Pearl (2000) describes the intuition behind causal power using counterfactuals. He says that causal power is the proportion of times that E did not occur when C was absent, but could have occurred if C were present. The numerator is just the raw increase in the probability of E when C is present, ΔP; the denominator is the proportion of cases under consideration—when C was absent and E did not occur, $[1 - p(\text{E}|\sim\text{C})]$.

Rehder's causal model theory of categorization uses the same mathematics underlying causal power. Specifically, causal model theory asserts that people's judgments of category membership reflect the joint likelihood of the item's observed properties, given one's knowledge of the category's generative causal relations. For example, consider a novel category of stars called "Myastars," which are defined by the features and causal relationships in Table 1.

If causal powers are probabilities, then these features and causal relations form a "fuzzy-or" network, in which one can compute the probability of each feature, F, being present according to Equation (1):

$$p(F \mid F_{\text{causes}}) = 1 - (1 - b_F) \prod_{C \in F_{\text{causes}}} (1 - m_{CF})^{C_{\text{present}}} , \qquad (1)$$

where F_{causes} is the set of direct causes of F, b_F is the probability that some "background" cause (e.g., a cause not mentioned in Table 1) leads to the presence of F, m_{CF} is the probability that feature C generates the presence of F (i.e., the causal power between nodes C and F), and C_{present} is an indicator variable equal to 1 when feature C is present and 0 otherwise. The equation reads:

feature *F* is present, given a value for each of its causes, when the disjunction of its various causal mechanisms fails to occur (the disjunction includes the background cause and the set of present candidate causes). The likelihood of a star with features *F1*, *F2*, *F3*, and *F4*, assuming it is a Myastar, is simply the product $p(F1)*p(F2)*p(F3)*p(F4)$.

Table 1. Features and relations of the artificial category "Myastars." Note: for ease of disposition, consider each feature to be "present/on" when it takes the value ionized helium, very hot, very dense, or many planets, and "absent/off" otherwise.

Features
(F1) Myastars are constructed from ionized helium
(F2) Myastars are very hot
(F3) Myastars are very dense
(F4) Myastars have a large number of planets
Causal relations
F1 causes F2: Ionized helium causes a star to be very hot.
F2 causes F3: Being very hot causes a star to be very dense.
F3 causes F4: Being very dense causes a star to have a large number of planets.

In a number of studies, Rehder (2003a, 2003b, Rehder & Kim, 2006) showed that people's ratings of category membership are well described by causal model theory (CMT). Rehder (2003b, Rehder & Kim, 2006) also distinguished CMT from the causal status view, finding that while causal status effects often do appear for "primary causes" (e.g., F1 above), they appear less robustly for downstream causes, such as F2 having greater influence than F3. CMT predicts less of a difference between the influence of F2 and F3, because these features have the same number of causes (just 1 each), and thus similar probabilities of occurrence.

CMT also predicts "coherence effects," whereby the presence or absence of two features jointly affects category judgments. In the causal chain example, an item with ~F1 and ~F2 is more likely than an item with F1 and ~F2, since ~F2 is actually more likely when ~F1. This distinction between CMT and causal status is reminiscent of the difference between "interactive"

and "independent" cue models of feature-based categorization, in which the features of an item contribute to the categorization decision with or without consideration of the other features, respectively. Exemplar models often outperform the prototype models (Medin & Schaffer, 1978; Nosofsky, 1986), and much of this distinction has to with the interactivity of the features (cf. Minda & Smith, 2002).

3.3 Causal Structure Learning and Intervention

The work of Ahn, Rehder and others (Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Waldmann & Hagmayer, 2006) shows that people's category-related judgments are based on knowledge of causal relations, but how are these causal relations learned in the first place? That is, how do people infer the presence and/or strength of a cause (or an entire network of causes) from their observations and interactions with category members? Until recently, most of the work on causal induction has addressed the learning of single causes based on contingency data, but new studies are investigating more complex causal networks and the role of interactive tasks in addition to contingency learning.

In one study, Steyvers et al. (2003) had people infer the causal structure underlying the contingencies between three events. Each event corresponded to the thoughts of a different "alien mind reader," and people were told that if one alien could read the mind of another, those aliens would share the same thought (a nonsense word; e.g., "TUS"). Participants were shown a series of trials in which the aliens thought of words and had the chance to read each other's minds, and then, were asked to choose which causal network represented their hypothesis about the mind-reading relations. Across different experiments, the number of possible networks raged from 2 to 18, where 18 is the number of all possible causal networks between three nodes with one cause

or two a-cyclic causal relations. (As one can see, the problem of structure learning is incredibly difficult when unconstrained, even with only three nodes.)

Steyvers et al. obtained a rich set of findings, but two are most relevant to this discussion. First, only a subset of individuals (~28% in Experiment 1) made optimal usage of the contingencies in choosing the correct causal structure, confirming the difficulty of the task. Second, when given the chance to "intervene" on the causal system (i.e., to fix the value of particular event, or alien thought), participants' beliefs about the underlying structure improved. In fact, it appeared that many of the interventions were chosen specifically to reduce uncertainty between equally plausible structures, given knowledge of just the contingencies. Thus, people can infer complex causal structures from a large set of alternatives (18 in this case), but only when given sufficient resources.

Other work anticipated the results of Steyvers et al., emphasizing the usefulness of interventions when learning a causal system (Pearl, 2000; Spirtes, Glymour, & Schienes, 1993). To demonstrate, consider the following example intervention. Imagine two possible causal structures between three nodes A, B, and C. In Structure 1, the underlying causes are A–>B–>C, and in Structure 2, A<–B–>C. In fact, Structures 1 and 2 are indistinguishable (also referred to as "Markov equivalent") given only contingency data, since one cannot determine purely from observations the direction of the causality between A and B. However, if one intervenes on the system by setting the value of B to be "present" (thus screening off the influence of A in Structure 1), then Structures 1 and 2 will yield different outcomes. In Structure 1, event B causes event C to be present but has no systematic effect on event A. In Structure 2, event B causes both C and A to be present. Clearly, intervention has benefits beyond that of mere observation.

Later work by Lagnado and Sloman (2004, 2006) showed that interventions like the two just described also improve learners' predictions. Those in an intervention group, but not those in an observation group, were sensitive to effects of "screening off," whereby two marginally dependent variables should be rated as independent when conditioned on a third variable. When the correct set of causes was Structure 1 from above, the intervention-based learners predicted that the probability of C was equal given B and given A&B. However, observation learners thought that adding A increased the probability of C, which is untrue.

Thus far, studies of causal structure learning have shown that inferring systems of causes with 3+ events is very difficult given only contingency data, but that certain tasks can improve learning, such as the ability to intervene on the causal system. The results are few but exciting, especially given the possibilities to extend this paradigm to cases of concept restructuring. The field is now ready to study the learning of more complex domains, armed with theoretical models, basic results, and new tasks that help learners go beyond association detection. The prospect of bridging work on conceptual change and causal learning is closer than ever.

3.4 A Paradigm for Causal Learning and Restructuring

The work on causal structure learning and intervention only brushes the surface of the next difficult problem for concept learning research: How do people infer entire systems of causes, going beyond just pairwise causal relations? Structure learning is evidently complex, but children and adults clearly do have knowledge of complex causal systems for real world domains. Thus, induction of causal systems is happening despite the difficulties, but how? The view developed in this thesis is that the learning of causal systems evolves incrementally over time (see also Fernbach & Sloman, 2009). By necessity, people learn subsets of causal structures piece by piece, and in the process, they are required to restructure previous causal beliefs. If this

is so, then people may struggle with learning entire systems of causes from scratch but show more proficiency in learning subsets at a time. Furthermore, models of causal inference might appear overly capable of complex structure learning when compared to all-at-once learning by people, but they may provide better fits for incremental learning shifts, and thus, help to identify the basic mechanisms underlying conceptual change.

This view has yet to be tested empirically. Current studies of structure learning assess the learning of entirely novel causal systems, but no studies have examined how people build upon and restructure their prior causal knowledge. Murphy's work has examined cases where prior concepts are invoked when learning later concepts, but in these studies the prior concepts are never restructured (Kaplan & Murphy, 2000). Work on order effects in causal induction suggests that what is learned from the first half of a set of contingency data may be overwritten by later contingencies (Marsh & Ahn, 2006), but the initial learning (and hence, what is restructured) is not typically evaluated. Fugelsang and colleagues (Fugelsang, Stein, Green, & Dunbar, 2004) examined the continual updating of causal hypotheses with the accumulation of data, but the initial hypotheses were only bolstered or reduced, not restructured. A developmental study by Schulz, Bonawitz, and Griffiths (2007) showed that 4 to 5-year old children inferred causal relations from evidence that ran contrary to their prior beliefs. However, their evidence for belief revision, as measured by transfer performance, was mixed. This study is perhaps the strongest empirical evidence linking studies of causal induction to concept restructuring.

Other current findings that do bear directly on concept restructuring are least tied to the formal theoretical approach. Chinn and Brewer (1993) documented the many ways that people react to anomalous data, only one of which (the least common) was genuine concept restructuring. Chinn & Brewer (2001) proposed a set of mental models for interpreting people's

verbal evaluations of anomalous data, but these were not formalized at the level specified in the causal induction models. Their work describes the reasoning that occurs in the background of a causal induction task—precisely what the models average across.

The goal of this research is to pick up where previous work drops off—namely, to develop and test a new paradigm to study the learning and subsequent restructuring of causal concepts, and to link this empirical work to the models of causal learning. As I have shown, such a paradigm is lacking outright in current work, and yet, the literature reviewed in cognitive development, science learning, and general learning in adults suggests that people's knowledge of causal systems plays a central role in concept learning and conceptual change.

To summarize, the studies to follow will use experimental and modeling techniques from the adult learning work to answer difficult questions posed by the conceptual change literature: How do people learn the structure of entire systems of causal relations? How might learning differ when done incrementally vs. all-at-once? And ultimately, what are the mechanisms underlying people's shifts in causal knowledge?

CHAPTER 4: EXPERIMENT 1—COMMON CAUSE

To begin, I developed a task in which individuals would learn and then restructure their hypothesized causal relations for a novel conceptual domain. The task was inspired by causal structure learning in real world domains, where one often develops a naïve, incorrect view of the underlying causal structure, and then with the accumulation of knowledge and evidence, restructures their original beliefs. Consider the structures provided in Figure 1.

Correlations between two events can be misleading, often suggesting a causal relation where none exists. This even occurs for professionals trained to resist causal attribution based on correlation alone. Rottman, Ahn, and Luhmann (in press) cite a famous example where an article published in *Nature* suggested that night-lights cause myopia in children, based on a correlation between these two factors (Quinn, Shin, Maguire, & Stone, 1999). A year after this publication (which received much publicity), another report suggested a compelling alternative account, based on two new correlations: (1) between incidents of myopia in parents and their children (i.e., heredity), and (2) between parents with myopia and the usage of nightlights, presumably due to poor vision (Gwiazda, Ong, Held, & Thorn, 2000). These new results suggest a "common cause" account of the original correlation—that parents with myopia were the cause of both factors. The previous belief in the direct cause was incorrect.

The common cause scenario is a quintessential way in which a learner may develop a prior, naïve concept and then need to restructure that concept based on new knowledge and evidence. Many other more complex changes may occur, but the common cause is perhaps the most fundamental. The first experiment takes the common cause as starting point to understanding the mechanisms underlying shifts in causal knowledge. Given that we currently know much about the initial learning of causal relations (i.e., the learning of the initial A causes

B link), this study asks how that initial learning affects the process of concept restructuring. In particular, how does the belief in the prior concept affect later learning where one views contingency data in favor of a different explanation? Do people revise their initial belief in the direct causal relation in favor of the common cause, and if so, how does this transition occur?

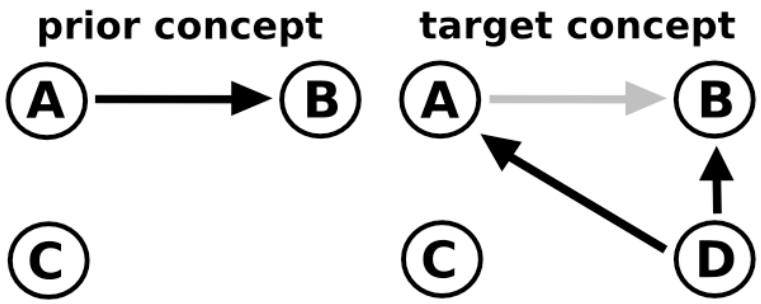

Figure 1: Diagrams of a hypothetical learner's causal representations en route to learning a common cause relation. The prior link remains in the target concept (the common cause), though reduced, signifying a possible residual belief in that link.

4.1 Predictions

Consider the possible effects the prior concept may have on inferring the target structure. First, the prior concept may serve as an anchor, or bias, such that people show commitment to the A.B link (A.B means "A causes B") and later learning of alternative causes is more difficult. Previous work shows that prior beliefs are difficult to give up, especially when they figure centrally in other causal explanations (for a review, see Koslowski, 1996).

Second, the acquisition of the prior belief may actually benefit later learning. In particular, evidence suggesting the lack of a correlation between other nodes in the system (between C&A and C&B) might draw resources away from those nodes and facilitate later search for the correct causal mechanism. This is especially true in Figure 1 since an alternative explanation for the A&B correlation is a mediating causal pathway, A.C.B. To the extent that

one can rule out this "mediating cause" explanation, they might consider the alternative common cause explanation sooner.

Third, both effects may occur. That is, learning the prior might increase one's belief in the A.B link, and independently, guide learners away from the wrong links and toward the right ones. If learners infer both the common cause and maintain a belief in the direct cause, they will have "over-explained" the occurrence of event B. Although previous work shows that people prefer simple explanations with fewer causal links (Lombrozo, 2007) and that competing causal hypotheses are considered in opposition (Lu et al., 2008), none have examined a case where learners are committed to a prior alternative conceptual structure, as is typically found in studies of conceptual change. In this case, people might over-explain to retain both possible causal pathways.

To test these predictions, I created an experimental paradigm analogous to Figure 1. One group, the "change" condition, was verbally instructed on a prior structure with three nodes where A directly causes B, then in a second phase, was shown a fourth node (D) and had to infer the correct causal structure from contingency data. The control group—the "no-change" condition—did not learn the prior structure and immediately attempted to infer the correct structure from contingency data with nodes A-D. The question is: How does the learning of the prior concept in the change condition affect the learning of the target concept, relative to that of the no-change condition?

Two dependent measures assessed learners' knowledge of the causal system. First, after the prior and target learning phases, participants rated the likelihood of each possible configuration of the system (all on/off combinations of the nodes, e.g., A/~B/C for the prior learning phase, A/~B/~C/D for the target learning phase). For each configuration, they were

asked, “How likely is the ecosystem to appear in the following way?” These judgments were used to infer participants “implicit” causal models of the system via model fitting, with the idea that some predictions offered above might not hold if participants were asked directly about their beliefs in the causal links (due to experimenter demands). Second, participants were asked at regular intervals during the target learning phase which of a set of possible links they believed were true. These judgments correspond to participants “explicit” beliefs about the causal system, similar to typical causal induction measures.

4.2 Method

4.2.1 Participants

Forty-eight University of Illinois students participated in exchange for course credit.

4.2.2 Materials

Participants learned about a fictitious ecosystem in the Pacific Ocean composed of four observable properties. Each property varied probabilistically during learning, taking one of two binary values (see Figure 2). The first property was the population size of a new type of fish biologists call “tespula”: above average or normal. The second property was the color of a new type of algae called “plemocyn”: very green or normal. The third property was the chemical composition of the barium contained in the ecosystem’s water: crystallized or not crystallized. The fourth property was the cloudiness of the water: cloudy or not cloudy. To simplify, I refer to the first values as the “on” values.

Figure 2. Properties of the fictitious ecosystem learned in Experiments 1 and 2. Items on top are the "on" values and items on bottom are the "off" values.

During covariation trials, the property values on a particular trial were determined by a causal system, like the one displayed in Figure 3 (the property roles were counterbalanced). The system was probabilistic, meaning that properties that are not caused may be "on" nevertheless due to a background cause not shown in the display (see Cheng, 1997). For the counterbalance condition depicted by Figure 3, the tespula population was more the average due to the background cause with probability 0.6. The tespula population being more than average would cause the barium to be crystallized with probability 0.85, and independently, the plemocyn to be green with probability 0.85. The water was cloudy due to a background cause with probability of 0.6. When the tespula population is average, all other properties would be "on" due to a background cause with probability 0.6. Background causes and observed causal mechanisms were combined using a noisy-OR gate, as shown in Equation 1 above.

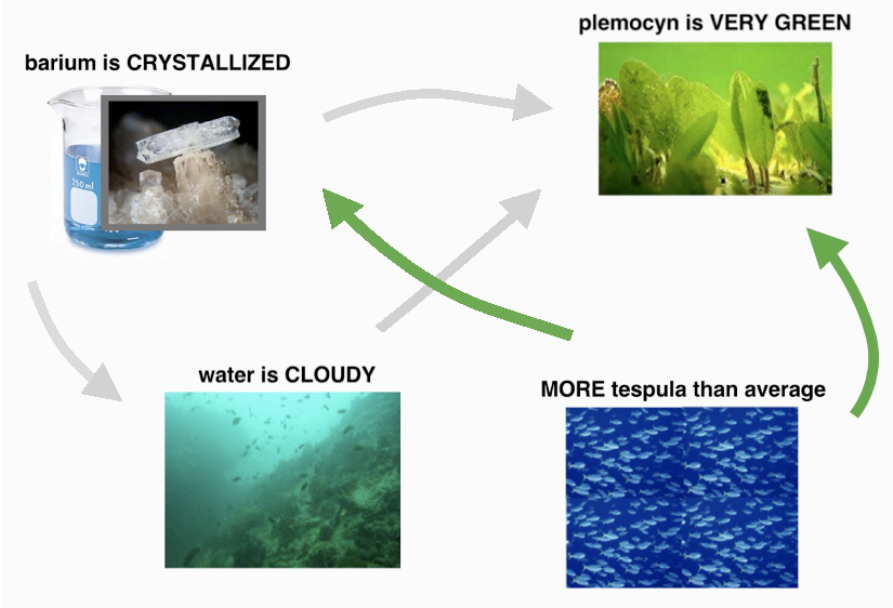

Figure 3. Example causal structure relating the properties of the ecosystem. Darkened links indicate that the properties share a generative causal relation in the indicated direction, with causal power equal to 0.85. All other potential links have causal power of 0. The properties will sometimes be referred to as A, B, C, and D, reading from left to right.

Covariation trials appeared as in Figure 3, except that the property values varied probabilistically and all arrows appeared in grey. During the test phases, participants were shown all "on/off" combinations of the four properties and told to rate their likelihood (see *Procedure* section). In this phase, the arrows were completely absent.

4.2.3 Design

Participants were divided into two conditions: "change" and "no-change," corresponding to those given a prior belief regarding the properties' causal relations before learning the target structure and those who learned only the target, respectively. Each condition was subdivided into four counterbalance conditions, controlling for which properties were assigned to the roles in the causal system. These conditions were determined by a Latin square, with the properties in the

following order: tespula population, barium crystallization, plemocyn color, and water cloudiness.

4.2.4 Procedure

The change condition is described first, since the no-change condition was a subset of the change condition procedures. Prior to the experiment, participants read and signed a consent form. Participants then read instructions and completed all tasks on a computer. For participants in the change condition, instructions stated that the purpose of the task was to learn about a new ecosystem in an obscure part of the Pacific Ocean. Their task was to help a group of biologists to understand how the properties of the ecosystems cause one another. Three of the properties of that ecosystem were described—the top two properties from Figure 3 (which I will call A and B) and the bottom left property (C). Participants were told that the biologists' current understanding was that property A causes property B (and told nothing else about C). They were also shown a picture with properties A-C and a green arrow connecting A to B. To ensure understanding, participants answered a multiple-choice question asking which properties were related and in what way. If they answered incorrectly, they repeated the instructions and re-answered the question until they were correct.

Next, participants entered the "prior learning phase" where they viewed a sequence of 30 "snapshots" of the ecosystem. Each snapshot depicted a particular on/off configuration of properties A-C. Each snapshot appeared with a frequency proportional to its likelihood, which was determined using the probabilities given in the Materials section and Equation 1. The snapshot frequencies were identical for all participants, but the order was random and different for each. Note that the causal system from Figure 3 creates a correlation between properties A and B, which in the absence of property D, supports the belief that A causes B.

After the 30 snapshots, participants entered the "prior likelihood rating phase" where they viewed each possible snapshot and were told to rate how likely the ecosystem is to look like the snapshot. Each item was rated twice, and then the ratings were averaged. They were also told, "when making the judgments, be sure to keep in mind the fact that the biologists think that [property A] causes [property B]." Ratings were given by moving a horizontal bar up and down a vertical scale, where the highest position indicated "VERY likely" and the lowest indicated "NOT likely."

Then, participants entered "the target learning phase." They were told that the biologists discovered an important new aspect of the ecosystem, property D, and now they are wondering if their previous belief that A causes B was "… wrong or perhaps missing something. For example, maybe [property D] causes some of the factors you were analyzing before."[2] They viewed a diagram similar to Figure 3 except with all links in grey, and were told their next task was to help the biologists figure out which of the portrayed potential causal relationships were true. Further, participants were told they would learn which causes were true by viewing snapshots like those in the prior learning phase. The instructions also clarified that each property may occur without being caused by another observed property (i.e., even if X causes Y, Y may appear in the absence of X) and that the links were not necessarily deterministic (e.g., if X causes Y, Y is simply more likely to appear in the presence of X).

Finally, they were told that in addition to viewing the snapshots, they would occasionally be making predictions about which of the causes are true; also, later during learning, the

[2] In real world learning, the discovery of a new cause often triggers new hypotheses implicating that cause. To ensure all learners inferred the same new causes (the common cause relations), I stated them explicitly. Admittedly, this may have biased the change condition's attention toward the common cause relations, unlike the instructions in the no-change condition. To ensure that this potential bias does not provide an alternative account for the results, I conducted an experiment identical to Experiment 1 but with a sixth possible causal relation from D to C. I also excluded the suggestion that perhaps D causes some of the factors from before. Participants in the change condition of this experiment chose links D.A and D.B 40% of the time before feedback, compared to 9% for link D.C, which was a significant difference, $t(19)=3.43$, $p<0.01$. Participants were not biased toward links D.A and D.B due to the instructions of Experiment 1 or to their novelty in the target learning phase.

computer would give feedback about whether their hypotheses were close to or far from the true structure. After every 10 snapshots participants were asked to guess which of the possible links were true. They were shown the picture in Figure 3 but with all links in grey, and told to click on the links to make their guess. When clicked, the links would darken.

To assist with learning, participants were given indirect feedback regarding their link choices starting on their 4th hypothesis trial (after 40 snapshots)[3]. They were never told the status of any particular link choice (e.g., that the A.B link was right or wrong). Instead, they were told that the hypothesis was VERY GOOD, GOOD, WEAK, or VERY WEAK, indicating that 5, 4, [3 or 2], [1 or 0] links were correct, respectively. Participants were not told the correspondences between the feedback and number of accurate links. On the final hypothesis, participants were told, "This is your LAST PREDICTION. On the next trial, make your best guess as to what causes what."

Finally, in the "target likelihood rating phase," participants again rated the likelihood of all possible snapshots of the ecosystem but this time with nodes A-D, each item twice. To recap, a flow diagram of the procedure for the change condition is in Figure 4.

The no-change condition was identical to the change condition, except that the prior learning phase and the prior likelihood ratings phase were excluded. The instructions immediately introduced participants to all four aspects of the ecosystem and the five possible links. Participants then began the target learning phase.

[3] Feedback was added to the learning task to improve learning. Previous work and the results of a pilot study show that learning is very poor for 3-4 node causal structures when given only covariation data (Steyvers et al., 2003, Lagnado & Sloman, 2004). Feedback is natural in real world learning and is usually provided by confirming or disconfirming predictions made on the basis of hypothesized causal relations. The feedback provided in this task can be viewed as a proxy for the outcome of multiple such predictions aggregated over time.

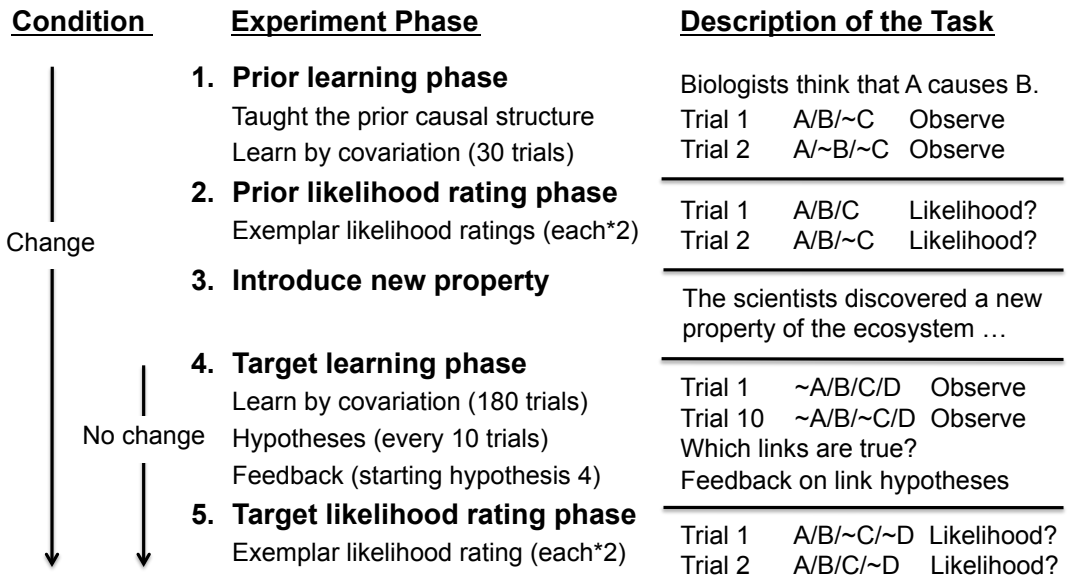

Figure 4. Flow of the experiment for the change and no change conditions. The change condition completed all phases, but the no change condition only completed phases 4 and 5.

4.3 Results

4.3.1 Hypotheses

First, I present the results from the hypotheses participants made during the target learning phase. Link A.B was analyzed separately, as was the pair of correct links D.A and D.B, and the pair of incorrect links A.C and C.B. (The of individual links within pairs were very similar). Hierarchical logistic regression was used to evaluate the effects of condition and hypothesis trial on link choice. The "hierarchical" component refers to a random intercept term, which was used to model the overall between-participant variability.

Results are plotted in Figure 5. To reduce inter-trial variability, I blocked the trials, except for the final trial: 1-4 (without feedback), 5-11 (feedback 1st half), 12-17 (feedback 2nd half), and 18 (the final trial). Main effects and interactions were assessed using likelihood ratio tests. These tests always compare a full model to a reduced model, where the full model includes

only the terms of interest (and in the case of interactions, the main effects as well.) The reduced model excludes only the terms being tested for significance.

The main effect of block on choosing the A.B link was significant, $\chi^2(1)=37.76$, $p<0.01$, suggesting that learning did occur, as participants selected this incorrect link less over time. The main effect of condition was not significant, $\chi^2(1)<1$. The interaction between block and condition was marginally significant, $\chi^2(1)=3.23$, $p=0.07$.

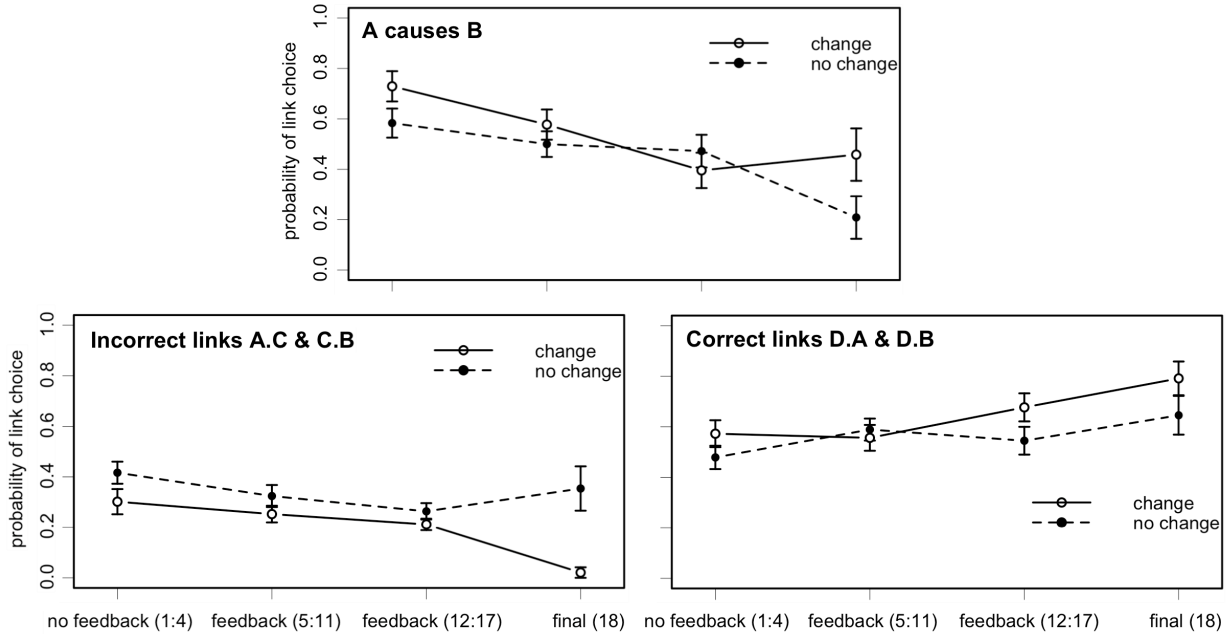

Figure 5. The probability of a participant including a link in their hypotheses during the target learning phase.

Because the difference in conditions for the A.B link was non-monotonic over blocks, two separate regressions were fit to trials 1-17 and trials 12-18. The interaction between trial and condition was significant for trials 1-17, $\chi^2(1)=9.13$, $p<0.01$, and for trials 12-18, $\chi^2(1)=6.77$, $p<0.01$, suggesting that the third block (i.e., trials 12-17) was different from early learning and

from the final hypothesis in different ways. Finally, the difference in conditions on just the final hypothesis was assessed using Fisher's exact test, which did not reach significance, $p>0.1$.

The interactions between trial and condition for the A.B link have two implications. First, although the change condition began selecting A.B more than the no-change condition, this difference went away by the third block as both conditions learned to not select A.B. Second, the difference in conditions increased from blocks 3 to 4. Relative to the no-change condition, the change condition was more likely to retain a belief in the prior concept in their final judgment, despite both conditions having chosen this link equally often during the final block of feedback.

The incorrect links A.C and C.B were analyzed together. The main effect of block was significant, $\chi^2(1)=23.02$, $p<0.01$, again showing learning. The main effect of condition was also significant, $\chi^2(1)=4.49$, $p<0.05$. The interaction between block and condition was not significant, $\chi^2(1)<1$. Although not originally predicted, this difference in the conditions is sensible, likely attributable to the extra learning in the change condition during the prior learning phase. The scientists' tentative theory regarding the ecosystem implied no causal relation between node C and either A or B. Furthermore, the 30 covariation trials suggested little correlation between these nodes, corroborating the scientists' view.

The correct links D.A and D.B were also analyzed together. The main effect of block was significant, $\chi^2(1)=15.74$, $p<0.01$. The main effect of condition was not significant, $\chi^2(1)=1.34$, $p>0.10$, though there was a tendency for to change condition to choose these links more often. The interaction between block and condition was not significant, $\chi^2(1)=2.63$, $p=0.10$.

In addition to analyzing each link separately, I conducted one final analysis to assess learning, this time in terms of whether the frequency of choosing only the two correct links (i.e., the correct explanation) increased over time. Results are plotted in the left panel of Figure 6. The

effect of block was significant, $\chi^2(1)=51.30$, $p<0.01$, but the main effect of condition was not significant, $\chi^2(1)<1$. The interaction between block and condition was significant, $\chi^2(1)=8.56$, $p<0.01$. The interaction reveals that despite the change condition being more likely to select the incorrect prior link, their choice in the correct two links increased more rapidly over time.

To compare, I also plotted the probability of choosing a combination of the correct links plus the prior, A.B. These appear in the right panel of Figure 6. Although the change condition is numerically above the no-change condition, neither of the main effects nor the interaction were significant, all $\chi^2(1)<1.5$. It appears that despite temptation to include the prior in one's hypotheses, the change condition is still more likely to choose the correct set of links. In other words, the prior learning phase helps more than it hurts.

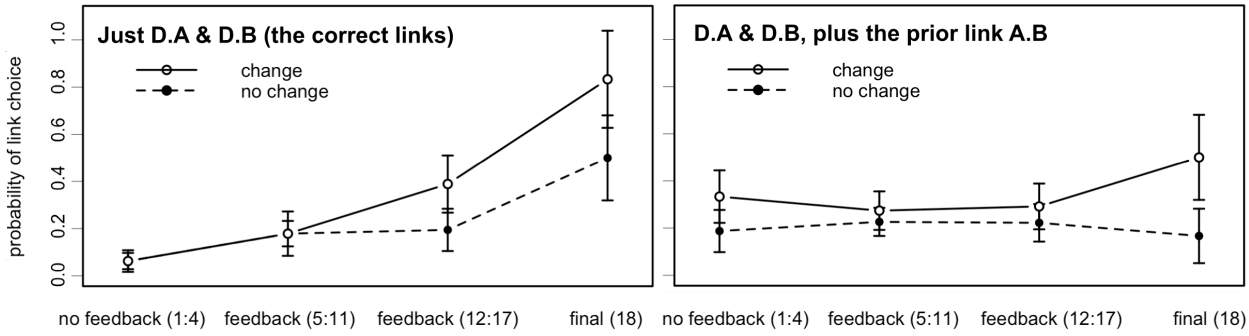

Figure 6. Probability of choosing only the correct links (left) or the correct links plus the prior link (right). To clarify, these selections always exclude both links A.C and C.B.

4.3.2 Likelihoods judgments

Likelihoods judgments were used to infer participants' latent causal representations via model fitting. To give a likelihood judgment the participant does not have to explicitly formulate judgments of causal strength, but the estimates are nonetheless driven by one's beliefs in the

underlying causes (Rehder, 2003b). Thus, we might expect differences in these two dependent measures if belief in some causes exists only at an implicit level; e.g., people may state verbally that they no longer believe in the prior, but an underlying belief may still guide their implicit judgments of the causal system.

Table 2. Average parameter fits (and standard deviations) from Experiment 1.

	No change	**Change**	***p*-values**
Background A	0.50 (0.08)	0.47 (0.09)	N.S.
Background B	0.42 (0.11)	0.36 (0.15)	N.S.
Background C	0.47 (0.06)	0.48 (0.04)	N.S.
Background D	0.53 (0.06)	0.55 (0.06)	N.S.
Link A.B	0.06 (0.08)	0.14 (0.14)	0.02
Link A.C	0.10 (0.11)	0.06 (0.08)	N.S.
Link C.B	0.07 (0.12)	0.05 (0.06)	N.S.
Link D.A	0.20 (0.15)	0.28 (0.16)	0.08
Link D.B	0.25 (0.16)	0.35 (0.20)	0.06
Links D.A & D.B	0.22 (0.13)	0.31 (0.14)	0.02

Causal model theory (CMT; Rehder, 2003a, 2003b) was fit to each individual's data using maximum likelihood estimation. Each fit yields an estimate of nine free parameters: the strength of each potential causal relation in Figure 3, plus an estimate of the causal background strengths for each node. The fitting routine worked by assuming that the participants' likelihood judgments were guesses about the relative frequency of the snapshots, should they be sampled again. Thus, 100 new snapshots were created with frequencies proportional to the normalized likelihood judgments of each participant. The MLE parameter values were those that maximized the likelihood of the snapshots.

The fits to CMT are presented in Table 2. Fitted background probabilities did not differ between the conditions, but estimates of causal strength were different, and in the same direction

as the differences present in the hypotheses data. First, the difference for link A.B was significant, $t(46)=2.39$, $p<0.05$, reinforcing the non-significant trend in the hypotheses data. This implies that the change condition represents the prior link stronger than the no-change condition, and this difference is robust for the more implicit measure, the likelihoods, where causal strength is not queried directly.

The conditions did not differ significantly in their representation of the incorrect links A.C and C.B, but the differences in the correct links were marginally significant: the change condition represented the D.A link more strongly, $t(46)=1.82$, $p=0.06$, as well as the D.B link, $t(46)=1.93$, $p=0.08$. Furthermore, when comparing the average of the two correct links D.A and D.B across conditions, the difference is significant, $t(46)=2.32$, $p=0.02$.

The latter result is in line with a predictions stated earlier that the change condition may benefit from the prior learning phase by observing the lack of a correlation between nodes A&C and between nodes C&B. Recall that links A.C and C.B constitute an alternative explanation of the A/B correlation; i.e., that A causes C causes B. Put simply, this set of links may be considered in opposition to the common cause links D.A and D.B in order to avoid over-explaining node B. If so, a reduced belief in the former may increase one's belief in the latter.

4.3.3 Inter-link inhibition

The idea that alternative causes compete or inhibit one another has empirical backing (Rehder & Milovanovic, 2007) and is made explicit in recent models of causal induction (Lu et al., 2008; Luhmann & Ahn, 2007). In the current study, to evaluate the relation between choices of links involving the two explanations, I used a hierarchical linear regression with number of correct links chosen as the dependent variable and number of incorrect links as the predictor. The predictor variable was separated into two parts: the participant-level effect (the average number

of A.C and C.B links chosen by a participant) and the within-participant effect (the number of links chosen on a given hypothesis minus the participant's average). These variables address different questions: the former asks whether participants who choose more incorrect links on average tend to choose more correct links; the latter asks whether on a given trial the number of incorrect links chosen affects the number of correct links chosen.

The effects of the two predictors were evaluated again via model comparison. A model excluding the between-participant effect did not fit worse than a model including both effects, $\chi^2(1)<1$. However, a model excluding the within-participants effect did fit worse than the model with both effects, $\chi^2(1)=52.03$, $p<0.01$, suggesting that causal links involved in competing explanations inhibit one another on a trial-by-trial basis. The model with only the within-participants effect, showed a negative relation between correct and incorrect links, $\beta=-0.26$ (standard error=0.04). To my knowledge this is the first evidence showing that competition in causal induction occurs at the level of entire explanations (i.e., sets of causes), beyond simply individual causal relations.

4.4 Discussion

The goal of Experiment 1 was to see how prior causal beliefs impact later learning and concept restructuring. Three possible outcomes were considered. One, the previously learned causal relation might impede learning of the alternative common cause relations if people are resistant to giving up their prior beliefs. Two, the previous learning may benefit later learning, since participants who learn the prior structure have extra evidence to help rule out the incorrect causal relations. Three, both effects may occur: People with prior learning may preserve the belief in the prior concept at an implicit level relative to the control condition, but they may also outperform the control in learning the correct structures. Results favored the third outcome.

Learners in the conceptual change condition "over-explained" property B relative to the control condition by maintaining belief in the A.B link in addition to learning the common cause links. These effects were obtained over a combination of hypothesis measures and model fits.

To recap, my account for these findings was that prior learning leads the change condition to focus less on the incorrect links during the target learning phase, and consequently, more on the correct links. Although the difference in conditions for the correct links was not robust for the hypotheses data shown in Figure 5, correlational analyses did support the tradeoff. The selection of links A.C and C.B was negatively associated with links D.A and D.B, and this effect occurred at the level of trials, not subjects, confirming the role of inter-link inhibition and not individual differences.

These findings present an interesting anomaly. On the one hand, both conditions treated competing causal explanations as inconsistent. But on the other hand, the change condition apparently did not treat common cause explanation D.A+D.B as inconsistent with the direct cause A.B, since they were more likely to select these three links in combination on their final hypothesis (though this difference was not significant; see Figure 6). It would appear then, that another difference between the change and no-change conditions is that change learners viewed their prior belief to be *less inconsistent* with the target explanation. This allowed them to maintain their belief in the prior in addition to learning the target structure; i.e., to "over-explain." If this finding were true in general (i.e., for other instances of conceptual change) it would have important implications for how educators should treat students' inconsistent prior beliefs. Namely, they should engage students in learning tasks that highlight the logical inconsistencies between the prior and the target concepts, to the extent they exist. This possibility will be tested in Experiment 3.

A few final analyses were conducted to verify this lowered perceived inconsistency in the change condition. Specifically, I determined which of the correct links (or both) was indeed viewed as less inconsistent with the prior belief. For each subject I computed how much more likely they were to choose A.B when NOT choosing D.A than when choosing D.A, and then the same difference for D.B. The mean differences are presented in Figure 7. The mean difference dependent on link D.A was 16.6% for the no-change condition and -1.4% for the change condition. The difference for the no-change condition was significantly greater than zero, $t(23)=2.98$, $p<0.01$, but the change condition difference was not, $t(20)<1$. The difference between conditions was marginally significant, $t(38)$[4]$=1.89$, $p=0.07$.

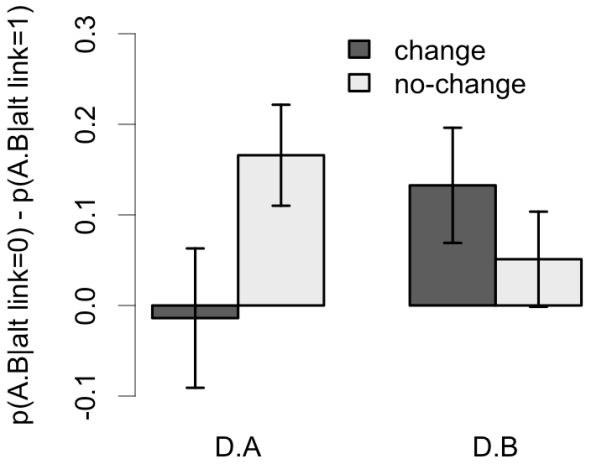

Figure 7. Differences in the probability of choosing the prior link A.B as a function of choosing links D.A or D.B.

The same comparisons were performed for the relation between links A.B and D.B. The mean difference in A.B choice dependent on D.B was 5.1% for the no-change condition and 13.3% for the change condition. The difference for the no-change condition did not differ from

[4] Degrees of freedom are off because this difference was not defined for all participants; some never chose D.A or always chose D.A. When conditions have unequal numbers of participants, t-tests used the Welch-Satterthwaite corrected degrees of freedom.

zero, $t(23)<1$, but the changed condition did, $t(23)=2.09$, $p<0.05$. The difference in conditions was not significant, $t(44)<1$.

I also tested the interactive effects of competitor link (D.A and D.B) and condition on choice of A.B, which was significant, $t(38)=2.63$, $p<0.05$. This interaction has two implications. One is that the difference between conditions (change showing less inhibition) was stronger for the D.A link. Hence, the D.A link may be the source of lowered inhibition, or perceived inconsistency, between the prior link and the correct links for the change group. Another implication is that link D.A was less inconsistent with the prior link than D.B for the change group compared to that of the no-change group. This may occur because D.A does not *directly* compete with the prior link A.B, since they lead to different effects. In contrast, link D.B does leads to the same effect as A.B, and current models predict competition between these links in particular (Lu et al., 2008; Rehder & Milovanovic, 2007). These accounts are both post-hoc, so I withhold further discussion until Experiments 2 and 3.

On the whole, results from the inter-link correlations suggest that for learners in the no-change condition, links that form competing explanations inhibit one another. This finding is consistent with previous work on explanation (Preston & Epley, 2009; Sloman, 1994) and current models of causal induction (Lu et al., 2008; Luhmann & Ahn, 2007; Rehder & Milovanovic, 2007). Going beyond previous work, this is the first study to show that link inhibition exists on a trial-by-trial basis when forming hypotheses about complex causal structures (i.e., with >3 nodes).

The results from the change condition are more nuanced, but also quite interesting, since they speak directly to the central goal of this thesis to identify the mechanisms underlying conceptual change, and resistance to concept restructuring. The change condition showed

relatively lower inhibition between the prior link, A.B, and part of the correct structure, link D.A. The prior learning phase altered the change condition's view of the inconsistency between the direct cause and the common cause. Later experiments will consider the role of this lowered inhibition in more detail.

4.4.1 Summary

The results from Experiment 1 were analyzed in greater depth throughout the discussion as I built a more complete story and fleshed out several implications. However, they come together to tell a coherent story. In general, learners with a prior concept of the causal system were more accurate in their hypotheses and representations of the correct and outright-incorrect relations. However, these learners were also resistant in giving up their prior belief, especially according to the more implicit measure, the likelihood judgments. This occurred despite the fact that the explicit measure, the hypotheses, did not differ between conditions at the end of learning.

Correlational analyses shed light on the mechanisms underlying these differences in causal learning. All learners tended to choose sets of causes in their hypotheses that formed coherent explanations and avoided selecting causes that were in competing explanations. Interestingly, however, this tendency was modulated for the change group, since these learners were more likely to select the common causes AND the prior, direct cause in the same hypothesis than the no-change group. Specifically, the change condition viewed the A.B link as less inconsistent with the D.A link than the no-change condition. This final result suggests that the change condition, in order to maintain their belief in the prior, treated elements of the target concept as less inconsistent with the prior concept.

CHAPTER 5: EXPERIMENT 2—CAUSAL CHAIN

The results from Experiment 1 suggest a number of exciting possibilities regarding the mechanisms of conceptual change. But given the novelty of the paradigm, replication is needed. The goal of Experiment 2 was to have participants learn a different target structure but with the same link possibilities from before. This will help to determine which effects from Experiment 1 are specific to the common cause scenario and which are generally true of causal restructuring.

Note, however, that the common cause is a quintessential form of causal restructuring. To wit, when deciding if two events share a causal relation, computing algorithms either assume there is no unobserved common cause leading to both events (Cheng, 1997) or attempt to infer whether there is (e.g., Scheines, Spirtes, Glymour, & Meek, 1994). Thus, learning scenarios like Experiment 1 where this assumption is later shown to be incorrect (when new events come to bear) are a prototypical source of conceptual change and may be distinct from other types.

There is one other way, however, that two events may be correlated and not actually share a direct causal relation. This alternative is depicted in Figure 3 by the combination of links A.C and C.B. If these two links are correct, then we call node C the "intermediate" cause, because it mediates the causal relation between A and B. In some cases, the need to clarify the intermediate cause might seem like an issue of pragmatics, since one can believe in both the direct causal relation and the role of the intermediate cause. For example, I strongly believe that drinking coffee causes increased productivity, even though I also understand that other known mechanisms mediate this relation: the transfer of caffeine to my blood, the consequent effect of stimulating my nervous system, and my compulsion to work when excited. At some level, it's both true that coffee causes hard work and that other causes mediate this relation.

Nonetheless, the intermediate cause has important consequences for causal reasoning. Although I insist on the direct relation between caffeine and work, I do admit that should any of the other intermediate mechanisms fail, that causal relation would be defective. Experiment 2 was a first attempt to discern whether people would learn intermediate causal relations after having learned a direct causal relation. To be consistent with Experiment 1, I gave feedback during the target learning phase suggesting that learners replace the direct cause with the intermediate causal relations (as Experiment 1). But whether the change condition will extinguish the prior causal relation in favor of the intermediate causal pathway, especially in light of the pragmatic issues, is an open question.

5.1 Predictions

Experiment 2 helps to test the generality of the many results from Experiment 1, including: (1) The change condition began selecting the prior link more than the no-change condition, but this difference disappeared later in learning (and then reappeared on the final hypothesis, though non-significantly); (2) The change condition selected the always-incorrect links (those excluding the prior) less often than the no-change condition; (3) Both groups selected the correct combination of links more often over time, but this selection increased more rapidly for the change condition; (4) The likelihood ratings showed that the change condition still represented the prior conceptual link more strongly than the no-change condition after covariation learning; (5) The likelihood ratings showed that the change condition represented the correct links more strongly than the no-change condition; (6) The selection of correct links correlated negatively with the selection of always-incorrect links, and this occurred at the level of trials, not individuals; (7) The change condition showed reduced inhibition (negative correlation) between the prior link and correct link D.A, relative to the no-change condition.

Which of these effects can we expect to obtain in Experiment 2? First, there is good reason to believe that effect (1) will replicate. As long as the change condition learns the prior link, they should begin selecting that link more, and then with feedback, select it less often. Effect (2) was originally unexpected but is sensible in retrospect. To the extent that prior learning is effective, the change condition should be less likely to select the incorrect prior links. Part of effect (3) should certainly hold—that people select the correct combination of links more over time—but whether the change condition shows faster learning will depend on the presence of effect (2) and inhibition between competing causal explanations. Effect (5) will likely appear under the same conditions.

Effect (4) is the current version of the standard finding in conceptual change research—learners with prior beliefs have difficulty giving them up, even when provided with evidence against them. If effect (4) holds, we have further evidence that the new paradigm succeeds in capturing this crucial aspect of conceptual change and lends confidence that other effects will generalize to real world cases of concept restructuring.

Effect (6) has support from previous work on explanation (Preston & Epley, 2009; Sloman, 1994) and causal induction (Lu et al., 2008; Luhmann & Ahn, 2007; Rehder & Milovanovic, 2007), and I expect it to replicate. The corollary, effect (7), was motivated by the contradictory findings that the change condition represented both the correct and the prior links more strongly than the no-change condition, and that alternative explanations competed for selection at the level of trials. If these two findings resurface in Experiment 2, I predict that at least one of the two correct links will show less competition between the prior for the change condition, relative to no-change. However, recall the issue of pragmatics, such that people may

find the direct cause and the causal chain more compatible than the direct cause and the common cause. Patterns of inhibition may change accordingly in Experiment 2.

5.2 Method

5.2.1 Participants and design

Fifty-six students from the University of Illinois participated in exchange for course credit. The design was identical to Experiment 1.

5.2.2 Materials

The physical materials were not changed in Experiment 2, but the probabilities associated with the causal graphs changed, as well as the properties displayed during the prior learning phase. In Experiment 2, the correct links were A.C and C.B, which together create a "causal chain" account of the correlation between A and B. Hence, during the prior learning phase, only properties A, B, and D were present to give the impression that A is a direct cause of B. This impression is created in the absence of the intermediate cause, just as in Experiment 1 the impression of A.B was created in the absence of the common cause.

The causal powers associated with these links were also changed. In a pilot study, I found that using 0.60 for all background causal strengths and 0.85 for the generative causal relations A.C and C.B led to poor learning of the prior concept A.B in the change condition. With these probabilities, learners inferred not only link A.B but also link D.B. To understand why, I examined the covariation trials given during the prior learning phase and discovered the reason. Consider the fact that the base rate of property B is very high, 0.87, given that it is the end of a causal chain. Now consider the marginal probability of events B and D occurring alone, p(D)*p(B)*[1-(p(A)], which is 0.87*0.60*(1-0.60)=0.21. That means that 6 trials contained only

properties D and B, which is equal to the number of trials with only properties A and B. Although a "rational" learner would use causal power, or minimally delta-p, to compute causal strengths, people often weight the trials in which both the candidate cause and effect occur (in this case D and B) more than all other trials (e.g., Kao & Wasserman, 1993). This weighting minimizes the trials that speak against the causal relation, such as the many trials in which B occurs but D does not occur.

To prevent the change condition from inferring an incorrect cause during the prior learning phase, I altered two of the parameters in the correct causal model. First, I lowered the causal strength from C to B to 0.45. Second, I lowered the background causal strength of D to 0.50. These have the combined effect of lowering the instances of just D and B to 4, whereas instances of just A and B are now 7. In the results from Experiment 2, I first present the model fits from the prior likelihoods to verify that the change in these probabilities led to the desired effect of limiting the causal inferences to just A.B and neither D.A nor D.B.

5.2.3 Procedure

Two minor changes in the instructions were implemented in Experiment 2, which was otherwise identical to Experiment 1. When I introduced the new property to the change condition the Experiment 1, I told participants, "Now the scientists are wondering if their previous belief that [A causes B] was wrong or perhaps missing something. For example, maybe [D] causes some of the factors you were analyzing before." This may have biased participants towards selecting links involving the new property D, and upon getting positive feedback they may have learned these links more readily than participants in the no-change condition. Although some evidence already speaks against this possibility (see footnote 4), the instructions in Experiment 2

were modified as follows, "[same] … For example, maybe [D] *or [C]* causes some of the factors your were analyzing before," (italics added to clarify the change).

The second change was very minor. The end of the instructions in Experiment 1 stated, "Pay close attention to the snapshots and try to figure out what causes what." In case these instructions were too open ended, in Experiment 2 I clarified by saying, "Pay close attention to *which snapshots tend to appear more than others* and try to figure out what causes what."

5.3 Results

5.3.1 Prior likelihood ratings (manipulation check)

CMT was fit to the prior likelihood ratings to determine which of the potential causal relations was inferred during the prior learning phase. In the previous version of Experiment 2, the fits to CMT revealed the following causal strengths: A.B=0.35, D.A=0.21, D.B=0.31. Only 16/21 participants showed a greater A.B causal strength than the average of D.A and D.B combined, and this proportion did not differ from chance ($p>0.1$, binomial test). In the current version of Experiment 2 with the modified causal model parameters, the causal strengths were A.B=0.46, D.A=0.25, D.B=0.18. In this case, 22/28 participants showed a greater A.B causal strength than the average of D.A and D.B ($p<0.01$, binomial test). Hence, the modified causal model (but not the model in the previous version of Experiment 2) had the desired effect of leading mostly to inferences about the prior causal relation A.B. This is crucial, since the prior learning phase was intended to establish a prior belief in *only* the A.B relation.

5.3.2 Hypotheses

The hypotheses data were analyzed as in Experiment 1. Results are plotted in Figure 8. For the prior link A.B, the main effect of block was significant, $\chi^2(1)=7.42$, $p<0.01$. The main

effect of condition was significant, $\chi^2(1)=7.14$, $p<0.01$, as was the interaction between block and condition, $\chi^2(1)=7.31$, $p<0.01$. The interaction indicates, as in Experiment 1, the change condition began selecting the prior link more than the no-change condition, but this difference decreased over time.

In contrast to Experiment 1, the change condition's preference for the A.B link did reduce to below the level of the no-change condition. To verify that conditions were not reliably different at the end of learning, I conducted an independent t-test on just the third block. In fact, during this block the change condition did not select the A.B link more than the no-change group, $t(53)=1.41$, $p>0.10$). Hence, by the final feedback phase, participants in both conditions believed in the prior link A.B to similar degrees.

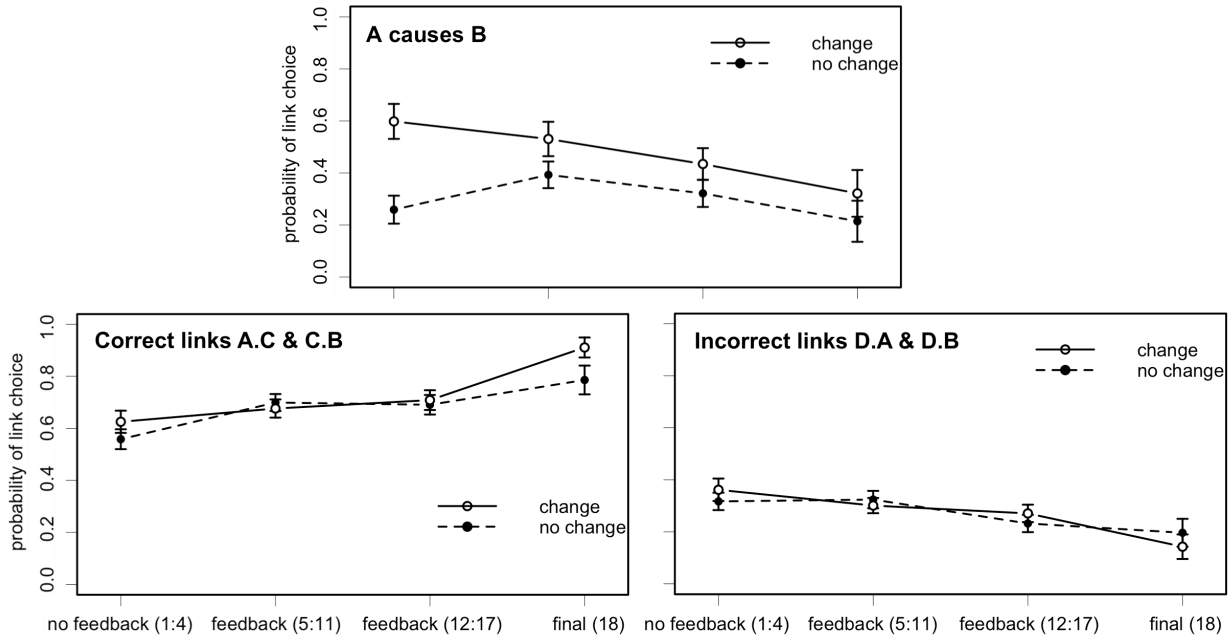

Figure 8. The probability of a participant including a link in their hypotheses during the target learning phase.

The regression on the incorrect links D.A and D.B revealed a significant effect of block, $\chi^2(1)=20.69$, $p<0.01$, but no effect of condition, $\chi^2(1)<1$, and no interaction between block and condition, $\chi^2(1)<1$. As the plot indicates, the difference in conditions on the final hypothesis is in the predicted direction but does not reach significance.

The regression on the correct links A.C and C.B revealed a significant effect of block, $\chi^2(1)=29.61$, $p<0.01$, but no effect of condition, $\chi^2(1)<1$, and no interaction between block and condition, $\chi^2(1)<1$. Although this overall lack of a difference in conditions for the correct links fails to replicate Experiment 1, there was a trend on the final hypothesis for the change group to select the correct links more than the no-change group (change=91% correct vs. no-change=79% correct, $t(45)=1.78$, $p=0.08$).

As in Experiment 1, I also assessed whether the frequency of choosing only the two correct links increased over time. Results are plotted in the left panel of Figure 9. By this measure, performance increased significantly over time, $\chi^2(1)=34.81$, $p<0.01$. However, the main effect of condition was not significant, $\chi^2(1)<1$, nor was the interaction between block and condition, $\chi^2(1)=2.17$, $p>0.10$. These findings mirror Experiment 1, except that the learning in the change condition did not increase at a higher rate than the no-change condition. This lack of a difference was expected given that the change condition did not select the incorrect links less often, and hence, were not biased toward the correct links as in Experiment 1.

I also plotted the probability of choosing a combination of the correct links plus the prior, just links A.C, C.B, and A.B. These appear in the right panel of Figure 9. The main effect of block was significant, $\chi^2(1)=6.84$, $p<0.01$. The change condition is numerically above the no-

change condition, as in Experiment 1, but this difference is not statistically reliable, $\chi^2(1)=1.42$, $p>0.10$. The interaction is also not significant $\chi^2(1)<1$.

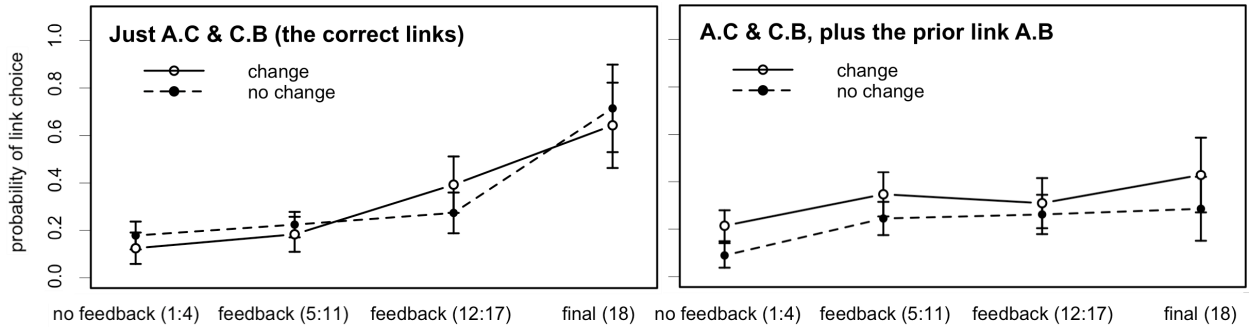

Figure 9. Probability of choosing the correct links (left) only or the correct links plus the prior link (right).

Overall, the individual link dynamics suggest two important similarities to Experiment 1, and few others. Critically, the change condition began selecting the prior link more often than the no change condition, and this difference became smaller over time. In addition, both conditions selected the correct combination of links more over time, showing that learning does occur for these complex structures when participants are given feedback on their link hypotheses.

5.3.3 Likelihood ratings

The likelihood ratings were fit to CMT as in Experiment 1, and the results are presented in Table 3. As in Experiment 1, the fitted background causal strengths did not differ between the conditions. In addition, the main result was replicated—the change condition represented the prior link more strongly than the no-change condition, despite the lack of a difference in the hypotheses in the final block. This difference suggests that participants' implicit causal representations show resistance in giving up the prior, relative to the hypotheses. Unlike Experiment 1, the change condition did not represent the correct links stronger than the no-

change condition. This lack of a difference was expected, however, given the finding that the change condition did not select the incorrect links less often. The potential advantage of increased attention to the correct links was absent.

Table 3. Average causal strengths (and standard deviations) from Experiment 2.

	No change	Change	*p*-values
Background A	0.50 (0.06)	0.50 (0.06)	N.S.
Background B	0.41 (0.11)	0.40 (0.17)	N.S.
Background C	0.50 (0.07)	0.47 (0.10)	N.S.
Background D	0.49 (0.05)	0.49 (0.05)	N.S.
Link A.B	0.08 (0.11)	0.20 (0.22)	0.01
Link A.C	0.25 (0.15)	0.24 (0.21)	N.S.
Link C.B	0.28 (0.21)	0.21 (0.18)	N.S.
Link D.A	0.10 (0.11)	0.11 (0.12)	N.S.
Link D.B	0.09 (0.10)	0.10 (0.13)	N.S.

5.3.4 Inter-link inhibition

Following Experiment 1, I assessed the inhibition between causes from competing explanations by predicting the number of correct links selected, this time A.C and C.B, with the number of incorrect links selected, D.A and D.B. As before, model comparisons were designed to tease apart whether inhibition occurred at the level of trials or individual learners. Experiment 2 replicated the findings from Experiment 1. A model excluding the between-participant effect did not fit worse than a model including both effects, $\chi^2(1)<1$, but a model excluding the within-participants effect did fit worse than the model with both effects, $\chi^2(1)=22.38$, $p<0.01$. In light of the apparent differences between the causal chain and the common cause scenario (i.e., the lack of replication in other results), this finding provides strong evidence that inter-link inhibition is a

stable mechanism underlying causal learning and restructuring. Causes that compete to explain the same property correlations inhibit one another at the level of individual hypotheses.

I also performed analyses to see if the conditions differed in whether they viewed the prior cause A.B and the correct causal explanation A.C+C.B as inconsistent. Recall, in Experiment 1 the change condition treated link A.B as inconsistent with the correct link D.B but not inconsistent with D.A, which might have enabled their persistent belief in the prior despite learning of the target links. So far, Experiment 2 replicated only some of the previous effects, leading to some uncertainty about further replications. In addition, there is the issue of pragmatics such that people may view the intermediate causal pathway as less inconsistent overall with A.B, given that these may reasonably co-exist as different levels of analysis.

To assess the inhibition between the prior cause and the correct causes, I computed for each subject how much more likely they were to choose A.B when not choosing A.C than when choosing A.C, and the same for link C.B. The mean differences are presented in Figure 10 (alongside results from Experiment 1, to compare). The mean difference dependent on link A.C was 5.7% for the no-change condition and 5.3% for the change condition. Neither difference was significantly greater than zero, both ts<1, and the difference in conditions was not significant, $t(52)<1$.

The difference in probability of choosing A.B dependent on the D.B link was -2.2% for the no-change condition and 11.7% for the change condition. The difference for the no-change condition was not significant $t(26)<1$, and the difference for the change condition was marginally significant, $t(24)=1.89$, $p=0.07$. The difference between conditions only approached significance, $t(49)=1.65$, $p=0.11$. Finally, the interaction between condition and link (A.C vs. C.B) was not significant, $t(49)=1.96$, $p=0.24$.

The inhibition results are different from Experiment 1 but interpretable. First, the no-change condition viewed neither link A.C nor C.B in strong competition with the prior link A.B. This is sensible if learners viewed the causal chain as complementary to the direct causal relation A.B. Second, in an interesting twist, the change condition showed a tendency to see the C.B link as inconsistent with A.B, but not the A.C link. This trend can be likened to the inhibition between links D.B and A.B from Experiment 1 (see Figure 10 to compare), since in both cases the causes that compete to explain the same effect, property B, showed the most inhibition.

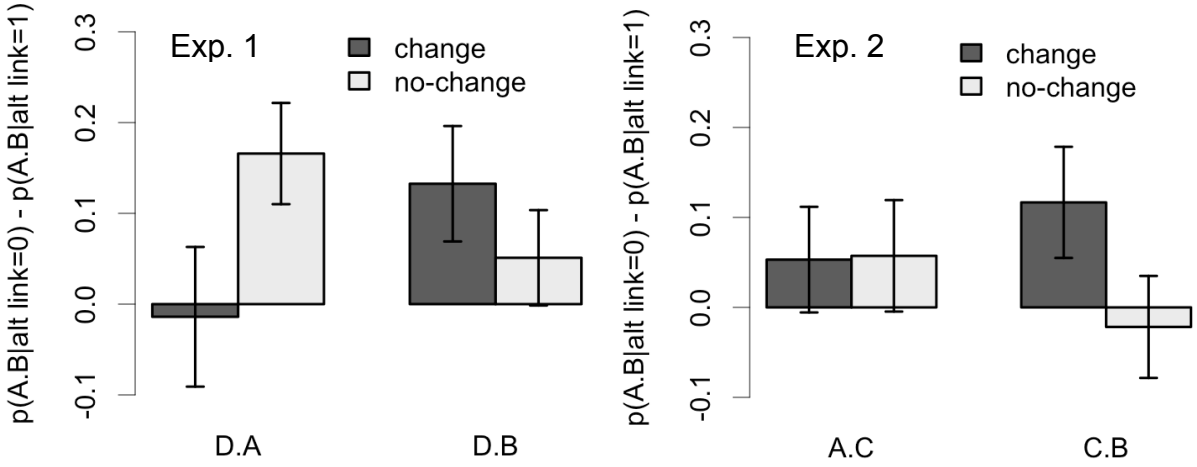

Figure 10. Differences in the probability of choosing the prior link A.B as a function of choosing links D.A or D.B (Experiment 1, on left) and A.C or C.B (Experiment 2, on right). Both experiments are plotted to facilitate comparison.

Taken together, these findings suggest that the change condition, in attempt to find a compromise between their prior concept and the target structure, may be fragmenting the target structure into "causal pieces" and more often choosing the piece that is not directly inconsistent with their prior concept. This fragmenting hypothesis makes a testable prediction—that encouraging the change condition to develop more complete, as opposed to fragmented, model of the causal system may help with learning. That is, if they view the prior concept as

inconsistent, on the whole, with the target concept, they might abandon it sooner in favor of the correct links. Future experiments will address this possibility.

5.4 Discussion

In summary, Experiment 2 provided three important replications in addition to shedding light on some interesting differences in how people learn common cause and causal chain relations when their prior knowledge suggests a direct causal relation. First, as in Experiment 1, learners in the change condition began selecting the prior cause more than the no-change condition, but this difference decreased over time. Second, the likelihood ratings indicated that belief in the prior cause was still present for most change learners, and different from the no-change learners, despite the continual decrease in hypothesizing this link over time. Third, all learners chose causal relations in their hypotheses to the exclusion of other causes that were involved in competing explanations.

The findings of Experiment 2 also differed in several ways from those of Experiment 1, but these differences are sensible and were consistent with predictions outlined above. The one difference that did not re-appear for which I cannot readily explain is the finding that change learners did not select the incorrect links less than the no-change learners. A possible post-hoc account is that despite the alterations in the causal model described in the methods section, the change learners did not learn to fully ignore the D.A and D.B relations during the prior learning phase. This account fits with the observation that the lowest mean causal strength inferred from the prior likelihood ratings in Experiment 2, D.B=0.18, is greater than the highest mean causal strength (among the two incorrect links) from Experiment 1, A.C=0.16. The prior learning phase may have been less effective for the change group in Experiment 2, and this could have much to do with the specific parameters used in the observation-generating causal model. Nonetheless,

given that the change condition did not have a bias against selecting the incorrect links, it makes sense that they did not show the corresponding bias toward the correct links. It makes further sense that they did not learn to choose the correct combination of links more rapidly than the no-change condition, as was found in Experiment 1.

The other differences, namely in patterns of inhibition between competing causes, can be explained by recalling the pragmatics issues with the causal chain and its hierarchical sets of causal relations. Today, perhaps more than ever, I believe that coffee causes me to work harder, and this belief is not the slightest bit compromised by my knowing that other lower-level biological and psychological mechanisms operate to produce this effect. Participants in both conditions showed little inhibition between the prior link A.B and link A.C, just as I would hypothesize that coffee causes me to work hard AND causes caffeine to enter my bloodstream.

Interestingly, however, the change condition showed modest inhibition between the prior link A.B and the competing link C.B. This effect resembles the inhibition from Experiment 1 between links A.B and D.B and suggests that conceptual changers are aware of (and acting in mind of) their belief in the prior link. To avoid over-explaining the effect property B, and to maintain their belief in the prior, this group preferred the link that was not a direct competitor with the prior link.

CHAPTER 6: INTERIM SUMMARY

Experiments 1 and 2 examined two prototypical forms of causal learning and restructuring—that of the common cause and the causal chain. One group of learners had prior causal knowledge that interfered directly with the target structure but the other did not. Comparing these groups was an important step in bridging studies of conceptual change with the basic research on concept learning. In both experiments, I showed that conceptual change effects appear even in these controlled experimental settings. Learners with a prior causal belief maintained that belief despite apparent learning of the correct alternative causes. This happened at the cost of "over-explaining," since the prior cause and the target causes were redundant. This finding runs counter to previous work showing that people prefer simpler explanations (Lombrozo, 2007), and thus, shows an importance difference in causal inference in cases with and without specific prior beliefs.

The purpose of bridging these two areas was to use the formal approach from basic research on concept learning to better understand the mechanism underlying conceptual change. Experiments 1 and 2 suggest two mechanisms. First, sets of causal links that formed coherent explanations inhibited one another. Specifically, the number of A.C and C.B links (together, the "causal chain" account) people selected on a given hypothesis restricted the number of D.A and D.B links (together, the "common cause" account). Said differently, learners built causal explanations consisting of multiple links, and when making hypotheses, they chose sets of links to the exclusion of other sets that compete to explain the same covariation statistics. This account extends previous work showing inhibition between individual causes that lead to the same effect (Lu et al., 2008; Luhmann & Ahn, 2007; Rehder & Milovanovic, 2007).

Second, a nuanced pattern of inhibition occurred between the correct links and people's prior conceptual knowledge, the A.B link. Among each pair of correct links, D.A+D.B in Experiment 1 and A.C+C.B in Experiment 2, one link is a direct competitor with the prior and one is an indirect competitor. Links D.B and C.B are direct competitors with the prior link A.B, because they lead to the same effect. Links D.A and A.C are indirect competitors, because they are inconsistent with the prior only when combined with their direct competitor link to explain the covariation between features A and B. Interestingly, learners in the change condition showed stronger inhibition between the prior and direct competitors. This finding suggests that learners undergoing conceptual change are biased toward accepting new causal links that do not compete directly with links from their prior knowledge. Learners appear to accommodate, rather than replace, their prior causal knowledge with new causal relations.

To put these results in context, recall that the goals for this thesis were twofold—first, to better understand the mechanisms underlying concept restructuring, and second, to use these insights to design better learning tasks that will promote conceptual change. Whereas the preceding experiments addressed the first goal, the final experiment will pursue the second goal, focusing just on just the change condition and on tasks that may assist in learning. Although many tasks have the potential to help with learning, I chose one that has improved learning in previous work and is especially suited to the learning of causal systems. The next chapter introduces this task, called "intervention," and presents an experiment to examine the role of intervention in concept learning and restructuring.

CHAPTER 7: EXPERIMENT 3—THE ROLE OF INTERVENTION

When learning about a causal system, it often helps to interact with the items that are presumed to cause one another. For example, students run experiments in physics labs to demonstrate and test the causal relations stipulated by physics theories. Experiments are literally "interventions" on the physical world, in which various entities are governed by underlying causal systems, like force causing acceleration. Psychological experiments are guided by the same intuitions. By intervening on the learning conditions in Experiments 1 and 2, I could tell how different conditions, such as having prior learning or not, affected future learning and concept formation.

The motivation behind Experiment 3 is that interventions may help learners attempting to restructure their causal knowledge. Furthermore, some types of interventions may be more helpful than others. The primary fault of learners from the change conditions of Experiments 1 and 2 was their tendency to maintain the prior belief in spite of apparent counterevidence. The hope for Experiment 3 is that interventions will prompt learners to challenge this prior and revise their concept in favor of the correct structure.

Much previous work has shown benefits of intervention compared to observation-based learning (e.g., Gopnik et al., 2004; Lagnado & Sloman, 2004, 2006; Steyvers et al., 2003). To see why, consider the sheer difficulty of learning the structures from Experiments 1 and 2 by observation. Although technically, the covariation trials are sufficient to learn the correct links (and rule out the incorrect links), the average learner has neither the foresight to compute the diagnostic conditional probabilities (i.e., causal power), nor the cognitive resources needed to track the numerous events frequencies that enter these computations. Interventions simplify the search for causal structure. One can intervene on a potential cause (e.g., by setting its value to

"on" or "off") and look directly for changes in the probability of the effect. If the effect varies for different values of the cause, the causal relation is supported.

Imagine being able to intervene on the common cause system from Experiment 1 where property D caused properties A and B. To test the prior relation, one has two options. First, intervene directly on property A and determine whether B is more likely when A is set to "on" than "off." Second, intervene on property D and observe the lack of a relation between A and B—both will be more likely when D is "on," but B no more likely when A is also "on" by chance. The former is more efficient, but both are simpler than passive observation.

Given the apparent benefits of intervention, the main prediction of Experiment 3 is that learners intervening on the ecosystem properties will find direct evidence against their prior belief, which will lead to more accurate hypotheses and likelihood ratings. Crucially, however, success will depend on the types of interventions learners tend to choose. As shown, interventions on A and D clarify the lack of a causal relation between properties A and B, but other interventions (on B and C) have no bearing on the A.B relation, at least no more than observation trials. Hence, the specific prediction of Experiment 3 is that learners choosing *helpful* interventions will outperform learners choosing less helpful interventions. In the next section, I describe an operational measure of intervention quality to validate the intuitions given by the A and D examples. Then, I show how this measure will be used to assess the relation between intervention quality and learning performance.

7.1 Defining Intervention Quality

Previous work describes good interventions as those that lead to greater confidence in the correct causal structure, relative to the incorrect structures (Steyvers et al., 2003). Formally speaking, the best interventions minimize uncertainty, or "entropy," defined over the set of

possible causal hypothesis; each hypothesis represents a set of possible links between properties A-D. Uncertainty, H, is captured by Equation (2):

$$\langle H(a)\rangle = -\sum_{y} P(y \mid a) \sum_{g} P(g \mid y,a) \log P(g \mid y,a), \tag{2}$$

where the expression in the inner summation $P(g|y,a)$ is the probability of a particular hypothesis, represented by "graph" g, given the intervention "action" a, and the resultant observations y. The outer summation takes this uncertainty for each possible outcome y of intervention a, weighted by its probability, $P(y|a)$, which yields the expected uncertainty $H(a)$. In Steyvers et al. (2003), $P(y|a)$ was computed by summing across all possible graphs obtaining the posterior probability of the outcome given the intervention,

$$P(y \mid a) = \sum_{g} P(y \mid g,a) P_s(g). \tag{3}$$

Using Bayes' rule, it can be shown that $\langle H(a)\rangle$ is computable knowing just the values of $P(y|g,a)$ and the subjective prior probability of each graph, $P_s(g)$, where the subscript s denotes that this is a subjective distribution corresponding to participants beliefs about which graphs are more likely than others (see Appendix for derivation). Values of $P(y|g,a)$ are computed by the generative model, g, with model parameters (i.e., causal strengths and background cause probabilities) equal to those of the generating model. Using a constant set of parameters simplifies the integration.

The other element needed to compute Equation (2) is $P_s(g)$, learners' beliefs about the prior probability of each graphical model. Selecting an appropriate prior distribution is crucial, because Equation (2) computes what interventions are most discriminating *if* their beliefs about the possible causes are $P_s(g)$. A natural choice for this prior distribution is the posterior probability of each graphical model, given the observed data. This is the prior distribution that

learners will calibrate to over time, if they take maximal advantage of the learning trials. To obtain the posteriors, I computed the normalized likelihood of each possible graph (of which there are 2^5–1=31, where the minus one eliminates the null graph, which was not possible), given a large set of data generated by the true graphical structure. Then, with $P_s(g)$ and $P(y|g,a)$ in hand, I computed $H(a)$ for each property A-D used in Experiment 1, separately for each value of those properties (i.e., "on" and "off"). These $H(a)$ were scaled appropriately and translated into choice probabilities according to the Luce choice axiom: $H(a)/\Sigma_i H(a_i)$. The resulting distribution in shown in left panel of Figure 11.

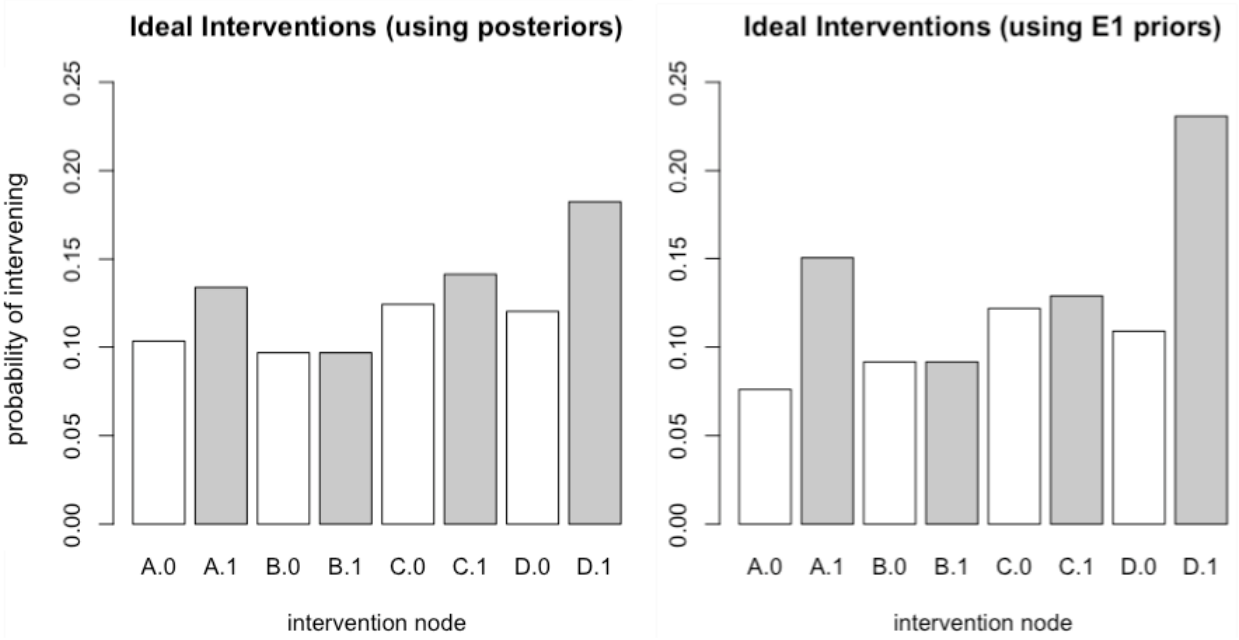

Figure 11. Ideal intervention probabilities for properties A-D (A.1 indicates setting property A to the "on" value, and A.0 means setting it to the "off" value.) This distribution of interventions will minimize uncertainty about the true graphical structure, if the prior graph probabilities that one is uncertain about are set by the posteriors given the observed data (left) or are equal to the probability that participants chose those graphs in Experiment 1 (right).

An "ideal" learner should select interventions with frequencies proportional to the probabilities in Figure 11. The exact values are determined by a scaling parameter, so only the

relative values are informative. First, note that turning properties "on" is more informative than "off," except for property B. This occurs because each graph makes a unique prediction only when its distinguishing set of causes is present. Otherwise, it predicts the same outcome as all other graphs without the distinguishing causes, and entropy (uncertainty) is highest when all graphs make the same prediction (i.e., have the same posterior probability). Property B is unique, since B is never a candidate cause.

Next, looking at just the "on" values, property D is the most informative. This is true because the observational data, and thus the posterior-based distribution $P_s(g)$, strongly favors models including relations from D. Also, D is a candidate cause of two other properties, so it is highly discriminating. Property A is informative for the same reasons, but less overall. A is also a candidate cause of two properties, but the posterior probability of these causal relations, A.B and A.C, are lower than the causes stemming from D. Interestingly, property C is also a good choice. Models with links A.C and C.B have a modest posterior probability, because they help to explain high base rate of B and also the correlation between A and B. Finally, property B should be intervened on least, because it is the only property that is not a candidate cause.

Overall, this "ideal learner" analysis of intervention quality fits with the intuitions given earlier about the usefulness of interventions on A and D. The one exception is that property C is more useful than anticipated, and is actually more useful than A. To test the generality of this result, I considered a different prior distribution, with the idea that a greater bias toward the A.B link in the prior (which is evidence by learners in the change condition in previous experiments) would lead to more emphasis on A interventions. The new prior distribution was taken directly from the hypotheses given by the change condition in Experiment 1—the probability of a given model was equal to the frequency that learners in the change condition chose that model, across

all hypothesis trials in Experiment 1, divided by the total number of hypotheses. This method had the desired effect of increasing the prior probability of models with the A.B relation from 0.15 (based on the posteriors) to 0.54.

The distribution of ideal interventions based on the new prior is presented in the right panel of Figure 11. As one can see, this adjustment leads to interventions on A being more helpful than interventions on C. Specifically, the prior probability of believing in link A.B appears to increase the informativeness of "on" choices for that link's cause. In the analyses to follow, I consider both intervention distributions when evaluating the role of intervention quality on learning. To foreshadow, the results do not depend on the choice of the prior.

7.2 Predictions

The general prediction for Experiment 3 is that learners conducting helpful interventions will make more accurate hypotheses, including more choices of the correct links and fewer choices of the incorrect prior link. The task will be identical to the change condition from Experiment 1, but in place of each observation trial, learners will choose a property to intervene on and observe the outcome. To test the effect of intervention quality, I will separate learners into two groups based on how closely their distribution of chosen interventions matches the ideal distribution show in Figure 11 (considering each ideal distribution separately), using the correlation, *r*, as the measure of closeness. Learners with above median *r* values should predict the correct links, D.A and D.B, more often, and the incorrect prior link A.B less often than learners below the median. Similar differences should arise in the CMT parameters.

To help distinguish alternative accounts for the effect of intervention quality, I will also have a second condition where, rather than choosing their own interventions, learners are told by the experimenter which interventions to make. Each participant in this condition will be "yoked"

to a participant in the standard intervention condition—they will be told to choose the exact sequence of interventions chosen by the learner in the standard condition (and hence, will push the same keys, to equate the sense of control). After each intervention they will also observe the same outcomes.

The yoked condition is important for two reasons. First, the median-split procedure creates a confound variable that may also lead to better performance in the above median condition—learners who choose better interventions may also be smarter or using more effort, which may also lead to better performance. This means I cannot evaluate the role of intervention quality, per se, with participants who generate their own interventions. Participants in the yoked condition, however, will be randomly assigned to their yoked counterpart. I will analyze this group separately to assess the effect of intervention quality.

The second reason for the yoked condition is to evaluate possible differences in learning by students selecting their own interventions compared to interventions assigned by a teacher (or the experimenter, in this case). Some previous work shows that learners benefit more from choosing their own interventions (Lagnado & Sloman, 2004; Sobel & Kushnir, 2006), because they are savvy to the organizing principles behind the specific intervention choices (Sobel & Kushnir, 2006). Going on previous findings, learners in the standard intervention condition should outperform those in the yoked condition.

However, the advantage to choosing one's own interventions is not guaranteed. Experiment 3 involves the learning *and revising* of previous causal knowledge, which differs in critical ways from the previous studies. Choice now entails a responsibility to use and interpret the interventions with an objective lens, but learners with prior beliefs often show a bias toward maintaining those beliefs despite counterevidence (e.g., Koslowski, 1996; D. Kuhn et al., 1988).

A compelling alternative prediction for Experiment 3 is that learners with a choice may interpret interventions on A as confirming the A.B relation, even if the outcomes are objectively disconfirming. (Probabilistic causes like those in Experiments 1 and 2 make this misperception even more likely.) In contrast, yoked learners may have a different, more critical, outlook on the intervention task colored by pedagogical demands. Or they may simply show less bias, since the intervention choices are not their own. In general, the role of intervention quality may depend on condition, and learners choosing their own interventions may retain a stronger belief in the prior relative to the yoked condition.

7.3 Method

7.3.1 Participants and design

Forty students from the University of Illinois participated in exchange for course credit. The design included two experimental conditions—the "standard" and "yoked" intervention conditions. Twenty participants were randomly assigned to each condition, except for the first few participants, who were in the standard condition by necessity.

7.3.2 Materials

Experiment 4 will use the same ecosystem materials from Experiments 1 and 2. The cover story for the interventions task is that scientists developed a model ecosystem to test different theories for how the ecosystem's properties relate. The scientists can selectively manipulate each property, and then observe whether and how it affects the other properties. An example was given for the method of manipulating the properties: "For example, [the scientists] can take crystallized barium from the ocean and put it into the tank to see whether the barium

crystallization affects the other aspects. They can also do this with the other aspects." The same example was given to all counterbalancing conditions.

The correct causal structure in Experiment 4 is Identical to Experiment 1—the common cause—but with different causal strengths and background cause probabilities to make the system more deterministic. Now, when property D is "on," A and B are "on" with probability 0.7, and the probability of all features being "on" due to some background cause is 0.25. The reason for changing to a more deterministic system was to ensure that participants could learn efficiently from a limited set of interventions.

7.3.3 Procedure

Both the standard intervention and yoked conditions performed the same task as the change condition from Experiment 1. The one procedural difference was that interventions replaced the observations during the target learning phase. Each condition performed 120 interventions, rather than 180 observations, to keep the task under 50 minutes. After every ten interventions, learners entered their current hypothesis about the causal system. Beginning the third hypothesis trial, learners received feedback on their hypotheses as in Experiments 1 and 2.

A learning trial for the self-generated condition proceeded as follows. First, participants selected a property of the ecosystem to intervene on. Second, they selected a particular value of that property ("on" or "off"). In the background, the program computed the probabilities of the various possible outcomes and used this distribution to sample one outcome at random. Third, after a short 1 second pause, participants observed the outcome selected by the computer program for 3 seconds.

The only difference in the yoked condition was that participants were told which property to intervene on and which value to set the property to. The instructions were similar to the

standard condition but rather than posing a choice, they indicated which specific property to intervene on. For example, instead of "Select an aspect to experiment with," the yoked condition was told, "Select the tespula population to experiment with." The same mouse clicks were required for both conditions.

7.4 Results and Discussion

7.4.1 Median-split conditions

In the first analysis, I separate learners in the standard intervention condition into those choosing more and less informative hypotheses. Each person's intervention distribution was compared to the graphs in Figure 11. A histogram of the *r* values are plotted in Figure 12, separately for the comparisons to the ideal interventions based on the posteriors (left) and on the Experiment 1 hypotheses (right).

Both histograms show a clear bimodal distribution, supporting the decision to split learners around the median. In addition, both groups suggest the same partitioning of learners, so it was not necessary to choose between the two prior distributions. To see how these groups differed, Figure 13 shows the averaged distribution of chosen interventions for the above median (left) and below median (right) groups. Participants above the median, as a whole, chose a pattern very close to ideal (compare with Figure 11). In contrast, participants below the median chose all properties with approximately equal frequency. Interestingly, both groups intervened the most on the A property—and the above median group much more than predicted—suggesting that all learners were attempting to test their prior knowledge of the A.B relation.

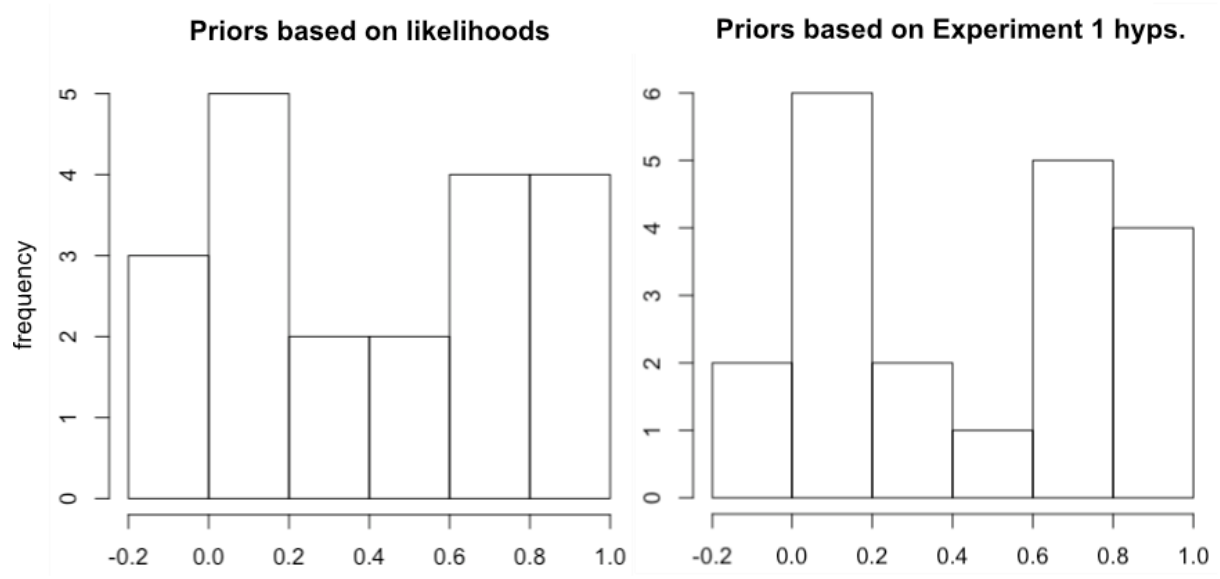

Figure 12. Histograms of the *r* values when comparing the interventions chosen by learners in the standard condition to the ideal distribution of interventions plotted in Figure 11.

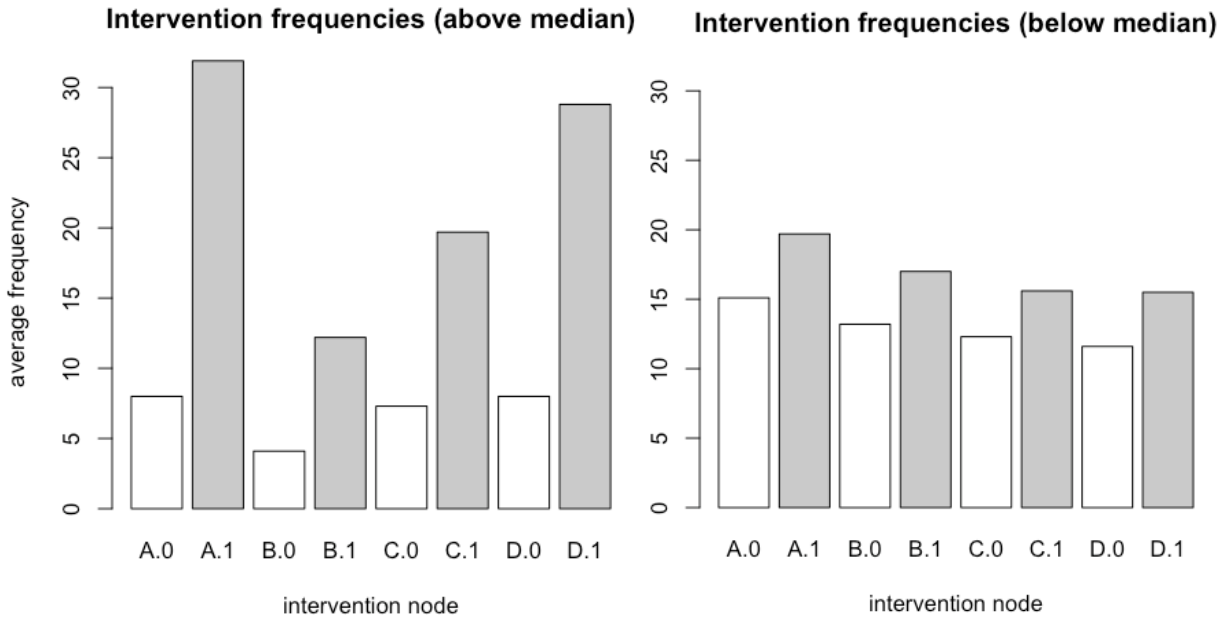

Figure 13. Average intervention frequencies for the above (left) and below (right) median split participants in the standard intervention group (and were also true of the yoked group.)

7.4.2 Hypotheses

To consolidate the number of analyses (there are now twice as many comparisons as in Experiments 1 and 2), I examine just two summary measures of learning—how often people chose the correct pair of links, D.A and D.B, and separately, how often they chose the correct pair of links plus the prior, A.B. These measures are a good proxy for overall performance, because they depend on all five links. The former requires learners to choose the correct links and none of the incorrect links. The latter requires learning for all links except the prior, and can be used to assess the degree of concept revision (or lack thereof).

I first compared the learning curves for the yoked participants with median split condition and hypothesis phase as a factor. The hypothesis phases were determined as follows: 1-3 (without feedback), 4-7 (feedback 1st half), 8-11 (feedback 2nd half), and 12 (the final trial). The probabilities of choosing just the correct links and the correct links plus the prior are displayed in the top half of Figure 14.

Looking just at the choice of correct links only, the main effect of block was significant, $\chi^2(1)=10.17$, $p<0.01$. The above median group is numerically above the below median group, but this difference was not statistically reliable, $\chi^2(1)=1.99$, $p=0.16$. The interaction was not significant $\chi^2(1)<1$. Looking at the choice of correct links plus the prior, the main effects and interaction were all non-significant, $\chi^2(1)<1$. Hence, although the above median group did choose the correct links more often, and included the incorrect prior link less often, these differences were not statistically reliable (though promising given the low sample size of 10 per condition).

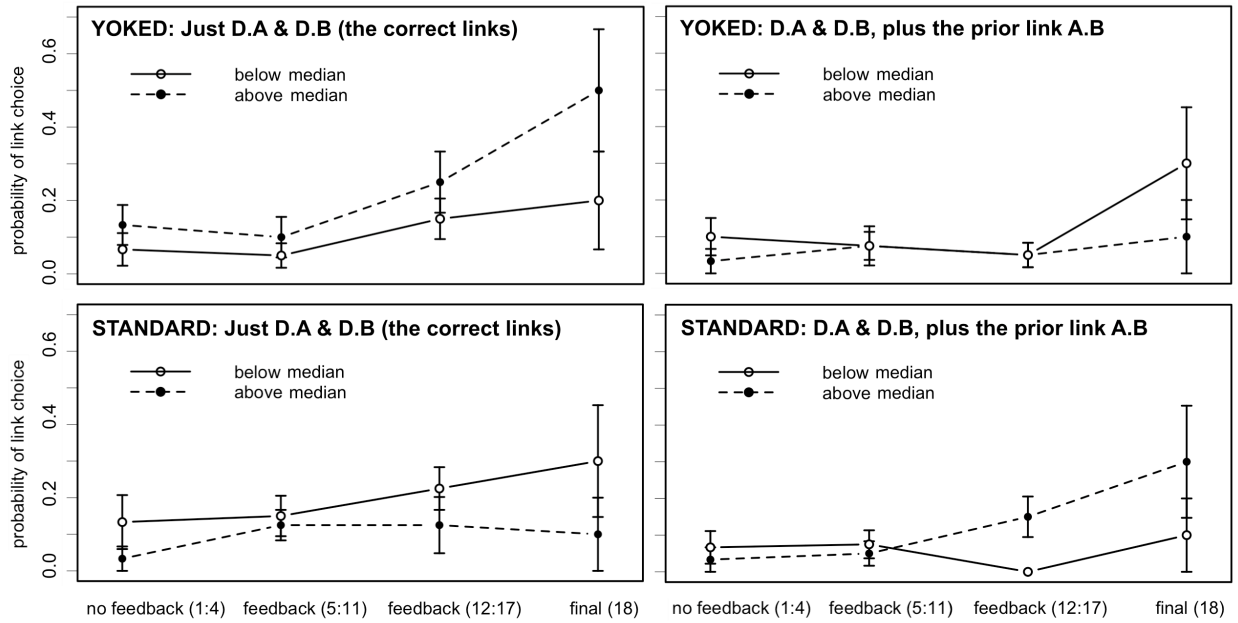

Figure 14. Probability of learners in the yoked (top) and standard intervention (bottom) conditions choosing the correct links (left) only or the correct links plus the prior link (right), as a function of median split condition.

I also compared the learning curves for the standard participants, again separating the above and below median groups to compare with the yoked participants. A very interesting pattern appears for the standard intervention group—in fact, the complete opposite pattern of their yoked counterparts. Participants who *chose* good interventions performed the worst overall, both compared to their yoked counterparts and those who chose poor interventions. This group was least likely to choose only the two correct links and most likely to choose the two correct links plus the prior. Looking at the choice of correct links only, the main effect of block was not significant, $\chi^2(1)=1.61$, $p>0.10$. The main effect of median split group was marginal, $\chi^2(1)=3.26$, $p=0.07$. The interaction was not significant $\chi^2(1)<1$.

Looking at choice of the correct links plus the prior, although the main effect of block was not significant, $\chi^2(1)=2.23$, $p>0.10$, the main effect of median split group was marginal,

$\chi^2(1)=3.00$, $p=0.08$, and the interaction was significant $\chi^2(1)=4.62$, $p<0.05$. Hence, learners choosing good interventions appeared to use them to their own detriment—they were less likely to choose only the correct two links and statistically more likely to increase their choice over time of the correct pair plus the prior, compared to learners choosing poor interventions.

These results suggest that the standard and yoked intervention conditions used the high-quality interventions in different ways—the latter to aid with learning the correct structure and revising the incorrect links, and the former perhaps to confirm their belief in the prior, leading to worse performance on both measures. To evaluate these opposing effects, I assessed the interactions between condition, block, and median split group. First, the 3-way interaction was not significant for choice of just the correct links, $\chi^2(1)<1$, but approached significance for the correct links plus the prior, $\chi^2(1)=2.60$, $p=0.11$.

For just the two correct links, the 2-way interactions were non-significant, except between intervention condition and median split group, which was marginal, $\chi^2(1)=3.54$, $p=0.06$. For the correct links plus the prior, the 2-way interaction between block and median split approached significance, $\chi^2(1)=2.30$, $p=0.13$, and again, the 2-way interaction between condition and median split group was marginal, $\chi^2(1)=3.05$, $p=0.08$.

Lastly, I analyzed responses on just the final hypothesis, given that all the differences between median split groups increased over learning. I consolidated the two measures of interest, coding each hypothesis as "correct" if only the correct links were chosen, "correct plus" for the correct links plus the prior, and "other" for all other hypotheses. Table 4 shows the breakdown for these responses distributed across the four conditions. To assess differences in the response distributions, I performed a multinomial regression, which is equivalent to performing two logistic regressions comparing "correct" vs. "correct plus" and "correct" vs. "other" as a function

of condition, median split group, and their interaction. The critical interaction between condition and median split group was a significant predictor of choice between "correct" and "correct plus," $\chi^2(1)=4.35$, $p<0.05$, as predicted, but was not a significant predictor of choice between "correct" and "other," $\chi^2(1)=1.93$, $p=0.17$.

Table 4. Frequency of each final response type across the four conditions.

	Correct	**Correct plus**	**Other**
Standard/above	1	3	6
Standard/below	3	1	6
Yoked/above	5	1	4
Yoked/below	2	3	5

Overall, participants' link hypotheses show dual effects of intervention quality. Learners *choosing* good quality interventions kept their prior belief in addition to learning the correct links, but learners merely *implementing* good quality interventions revised their belief in the prior and selected only the correct links. Earlier, I suggested that learners with a choice might interpret the outcomes of their interventions differently, given previous work showing that people have a bias to maintain their prior beliefs (e.g., Koslowski, 1996; D. Kuhn et al., 1988) and to collect evidence that reinforces this bias (Wason, 1960). In this study the standard and yoked conditions made the same interventions and viewed the same outcomes, so the bias was purely cognitive in nature, no more reinforced by the intervention choices for one condition than the other. Unfortunately, given the design of Experiment 3, it may be impossible to know for sure the intention of learners who chose their own interventions; e.g., whether they were looking to confirm their prior beliefs or challenge them. To discover the purpose of learners' interventions will be the focus of future research.

7.4.3 Likelihood ratings

As before, the likelihood ratings were fit to CMT. One participant was removed from the yoked condition, because they gave the same likelihood rating for all items. The results are presented in Table 5, separating the ratings from the standard and yoked conditions to see if the type of intervention affected the link representations. The fitted background causal strengths did not differ between the conditions, but interestingly, the strength of the A.B link was significantly greater for the standard condition than the yoked condition. None of the other links differed between conditions.

To save space and facilitate comparison, I have not further separated the model fits into the above/below median split conditions. None of the differences between median split conditions were significant, but to compare with the findings from the hypotheses, the A.B link was again higher for the above median group in the standard condition (0.19 vs. 0.12) and lower for the above median group in the yoked condition (0.7 vs. 0.8), though the latter difference may be muddled by floor effects.

Table 5. Average causal strengths (and standard deviations) from Experiment 2.

	Standard	**Yoked**	***p*-values**
Background A	0.48 (0.07)	0.46 (0.05)	N.S.
Background B	0.38 (0.12)	0.39 (0.10)	N.S.
Background C	0.47 (0.06)	0.46 (0.04)	N.S.
Background D	0.52 (0.05)	0.52 (0.05)	N.S.
Link A.B	0.15 (0.15)	0.07 (0.08)	0.04
Link A.C	0.03 (0.05)	0.06 (0.06)	N.S.
Link C.B	0.10 (0.14)	0.05 (0.07)	N.S.
Link D.A	0.15 (0.16)	0.18 (0.11)	N.S.
Link D.B	0.32 (0.18)	0.27 (0.17)	N.S.

The main effect of intervention type on link A.B complements the findings from the hypotheses. Learners choosing their own interventions seem to approach the task differently than yoked learners viewing the same trials. Choosing the interventions leads to a bias to confirm the prior relation, consistent with previous work on conceptual change and confirmation bias. The current findings add to this work by showing that biases persist even in an active learning task where learners produce direct evidence against the prior belief.

7.4.4 Inter-link inhibition

As before, I measured the inhibition between causes from competing explanations by predicting the number of correct links selected, D.A and D.B, with the number of incorrect links selected, A.C and C.B. Experiment 3 replicates the earlier findings. A model excluding the between-participant effect did not fit worse than a model including both effects, $\chi^2(1)<1$, but a model excluding the within-participants effect did fit worse than the model with both effects, $\chi^2(1)=7.22$, $p<0.01$.

I also determined whether learners viewed the prior cause A.B and the correct causal explanation D.A+D.B as inconsistent. Experiments 1 and 2 showed that learners who maintained their prior belief showed reduced inhibition between the correct links and the prior, specifically for the link that is not in direct competition with the prior (in this case, link D.A). Thus, given the results from the hypotheses, the above-median learners in the standard intervention condition should show reduced inhibition between link D.A and the prior link A.B. This difference should be absent, or perhaps reversed for the yoked condition.

The mean differences are presented in Figure 15. For the standard condition, the mean difference dependent on link D.A was -10.4% for the above-median learners and 16.7% for the below-median learners. The difference was not significantly different from zero for the above-

median learners, $t(9)<1$, and only approached significance for the below-median learners, $t(9)=1.66$, $p=0.13$. The difference between conditions was not significant, $t(17)=1.61$, $p=0.13$.

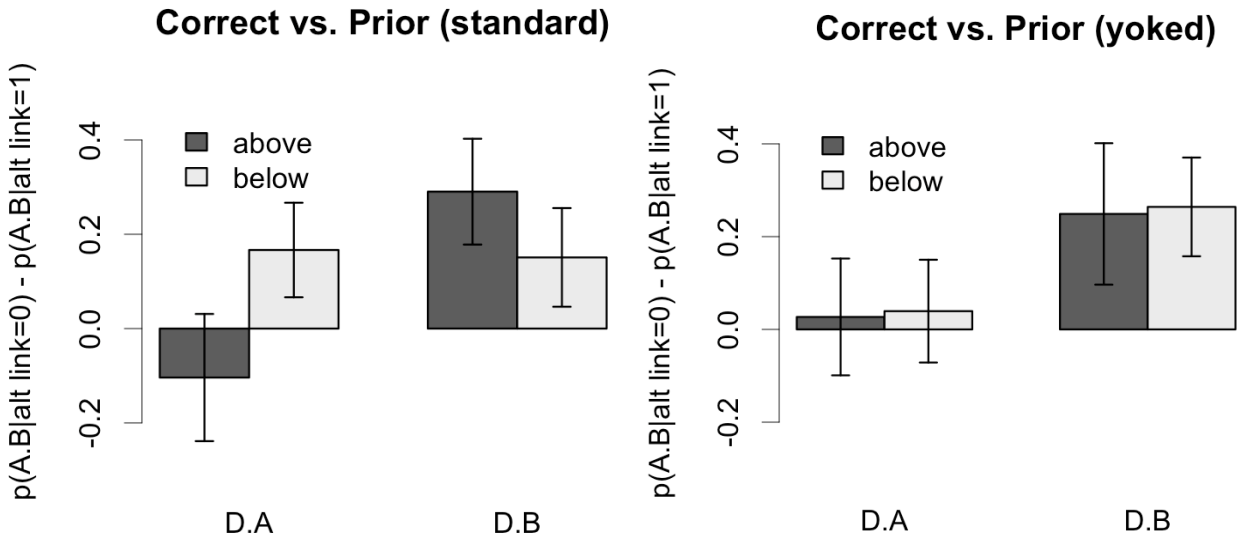

Figure 15. Differences in the probability of choosing the prior link A.B as a function of choosing links D.A or D.B, separately for the standard and yoked intervention conditions.

Again for the standard condition, the mean difference dependent on link D.B was 29.0% for the above-median learners and 15.1% for the below-median learners. The difference was significantly greater than zero for the above-median learners, $t(8)=2.59$, $p<0.05$, and was not significant for the below-median learners, $t(9)=1.44$, $p>0.15$. The difference between conditions was not significant, $t(17)<1$. Finally, as predicted, the interaction between condition and link (D.A vs. D.B) was significant, $t(15)=2.92$, $p<0.05$.

The yoked condition differed from the standard condition. Though the overall differences between links D.A and D.B were similar, there was no interaction between link and median-split condition. The mean difference dependent on link D.A was 2.7% for above-median learners and 4.0% for the below-median learners. Neither difference was significantly greater than zero, both $ts<1$, and the difference between conditions was not significant, $t(17)<1$.

Additionally, the mean difference dependent on link D.B for the yoked condition was 24.9% for the above-median condition and 26.4% for the below-median condition. The difference was not significantly greater than zero for the above-median condition, $t(8)=1.63$, $p=0.14$, but it was significant for the below-median condition, $t(9)=2.48$, $p<0.05$. The difference between condition was not significant, $t(15)<1$. The interaction between condition and link was not significant, $t(10)<1$.

Overall, the levels of inhibition between the prior and the correct links were as predicted. For the standard intervention condition, the above-median learners showed more inhibition between A.B and the directly inconsistent link, D.B, than between A.B and D.A. In contrast, the below-median learners (who also showed less commitment to the prior A.B link) had no preference for D.A or D.B. For the yoked condition, all learners showed more inhibition between A.B and D.B, with no effect of the median split conditions.

Altogether, Experiment 3 helps to build a more complete picture of the concept restructuring process. Interactive tasks can help with learning, but this effect depends on the quality of the interactions and one's perspective on the task. Further implications are apparent for classroom learning and other pedagogical settings. Connections to the literature on active learning and ideas for future work are discussed in the following section.

CHAPTER 8: GENERAL DISCUSSION

Learning is an ongoing, constructive process. People gather new ideas and evidence that bear on their prior knowledge, and in many cases that prior knowledge needs to be revised. Despite this truth, most studies on concept learning have addressed entirely novel concepts and cases where prior knowledge is minimally or indirectly involved. Studies of conceptual change address the role of prior knowledge head-on, but focus on a very different level of analysis, more descriptive and less concerned with the underlying mechanisms. A compete understanding of concept development requires both approaches, but few studies have addressed this unification.

8.1 Summary of Experiments 1-3

The goal for Experiments 1-3 was to unify these two approaches, building on recent work under the *theory view* of concepts (Gopnik et al., 2004; Murphy & Medin, 1985; Wellman & Gelman, 1992). Recently, theory theorists have argued that concepts are defined over networks of causal knowledge. The tenets of this view have been formalized using causal Bayesian networks, which have begun to shed light on the mechanisms of concept restructuring (e.g., (Griffiths & Tenenbaum, 2009; Kemp & Tenenbaum, 2008; Rehder, 2003b). The theory view is also adopted in prominent views of conceptual change (Carey, 1985; Vosniadou, 2008), and even alleged competitor theories agree on the crucial role of causal knowledge (e.g., diSessa & Sherin, 1998). Hence, studies of concept learning and conceptual change find common ground by emphasizing the role of causal knowledge.

8.1.1 Experiment 1

Experiment 1 advanced previous work on causal learning using a new task where learners developed a prior concept for a new causal system and were then prompted to revise it with new

evidence. Participants learned about a new ecosystem defined by a set of causal relations between four properties. The design exploited a common source of confusion when learning causal relations—the "common cause"—where two properties A and B are each caused by a third property D. In the absence of the common cause (property D) one might infer a causal relation between A and B, because they were correlated. However, upon viewing D and its relations to A and B, learners should replace their belief in A.B for the correct relations, D.A and D.B. The question was whether and how this change occurred.

To assess the role of prior knowledge when learning a causal system—i.e., what separates concept restructuring from novel concept learning—I compared a group who first learned the prior concept (that A causes B) and then viewed evidence that D was the cause of both, to a group that learned the entire system, including property D, from the beginning. The primary task for both groups was to observe a series of "snapshots" of the ecosystem, consisting of the four properties (A, B, D, and a distracter, C), generated probabilistically according to the common cause structure. I measured their knowledge of the causal relations *explicitly* by asking them what causes they thought were true at regular intervals, and *implicitly* at the end of the task, using a likelihoods rating procedure (see Rehder, 2003a, 2003b).

Three outcomes were possible based on previous findings. First, learners who received the prior concept first (the "change" condition) may hesitate to revise it, and thus, show poor learning of the correct structure. Second, learners in the change condition may benefit from the prior learning phase, which showed evidence for the lack of relations between properties A and B and the distracter property C. By knowing that C is not involved, they may focus more on the common cause, property D, during the primary learning task. Third, a mixture of these outcomes may occur. People may retain their belief in the prior, consistent with much previous work (e.g.,

Koslowski, 1996; D. Kuhn et al., 1988), but they may also show better learning of the correct links due to the de-emphasis on property C. Interestingly, this outcome would run counter to Lombrozo's (2007) finding that people prefer simple explanations, since people in the change condition may believe in both the direct cause between A and B and the common cause—in fact, either of these is sufficient to explain the correlation between A and B.

Results from Experiment 1 support the third prediction. Learners in the change condition had a stronger belief in the prior relation A.B at the end of learning, but they also outperformed the no-change condition in learning the correct relations D.A and D.B. Interestingly, the difference in belief for the prior relation A.B was robust only according to the implicit measure; both conditions were equally likely to select this relation in the final block of learning.

Results also revealed the mechanisms of causal learning and concept restructuring. For one, learners built hypotheses that formed coherent sets of causes and chose particular sets to the exclusion of others that explained the same covariation data. That means when learners chose links D.A and/or D.B, they were significantly less likely to choose links A.C and/or C.B, because both pairs explain the correlation between A and B. Conversely, when they did *not* choose A.C and/or C.B, they were free to choose D.A and/or D.B.

In principle, this inter-explanation competition should also apply between the prior cause A.B and the correct causes D.A and D.B. But in fact, the pattern was different for the prior. Learners in the change condition did show competition between A.B and D.B, presumably because they compete *directly* to explain property B, but showed no competition between A.B and D.A. The no-change condition was, if anything, more likely to select A.B along with D.A, suggesting that learners in the change condition were especially aware of the conflict between

the prior concept the correct structure. When keeping their prior belief in A.B they were more likely to accept only the correct relation that was not a direct competitor with the prior.

8.1.2 Experiment 2

The second experiment was a replication of the first, using the same set of causes, but a different target structure. The true causal relations were A.C and C.B, which form a causal chain between A and B, and thus, also create a correlation between A and B. As before, the change condition learned first that A causes B, and then later upon viewing property C was prompted to revise their prior belief.

Three crucial patterns replicated from Experiment 1. First, the change condition showed a stronger belief in the prior relation A.B at the end of learning, and again, only according to the implicit measure. Second, patterns of competition were evident between causal explanations. Choice of A.C and/or C.B correlated negatively with choice of D.A and/or D.B. Third, for the change condition, the prior relation A.B was in competition with only the directly inconsistent cause, now C.B. Interestingly, the no-change condition showed less competition overall between the prior and in the correct relations. This finding is consistent with a pragmatic view of the causal chain that the A.B relation is somewhat true at a higher level of analysis, even though it is mediated by property C.

Two findings did not replicate from Experiment 1, but these null results were sensible given post-hoc analyses. First, the change group did not choose the incorrect links less overall. Second, the change group did not show a greater belief in the correct links at the end of learning. A probable account for the former result is that the prior learning phase was simply less effective in reducing belief in the distractor causes. In support of this account, the likelihood ratings taken immediately after this phase show that learners still had a moderate belief in the distractor

causes, and more so than Experiment 1. Given the change group was not biased away from the incorrect links, it also makes sense that they were not biased toward the correct links. Taken together, this pair of null results actually supports their interpretation from Experiment 1. To the *extent* that learners can ignore the incorrect causes (in some cases, due to prior learning), they will better learn the correct causes. Otherwise, there should be no difference.

8.1.3 Experiment 3

The goal of Experiment 3 was to see if interventions would help with concept learning and restructuring, and also, if some types of intervention were more helpful than others. First, I tested the effect of intervention quality on learning, with the idea that interventions on critical causes—the correct causes and the prior cause—would be the most helpful. I divided learners in to two groups according to how similar their sets of interventions were to those of an "ideal" learner. Further, to avoid possible confounds introduced by the dividing procedure, I randomly assigned a second group of learners to sets of "yoked" interventions—each yoked learner was told to make the same interventions as a learner in the standard condition. I compared the learning performance of the good (close to ideal) and bad quality yoked learners. As predicted, the yoked learners with good interventions gave more accurate hypotheses than yoked learners with bad interventions (though this only approached significance, likely due to low sample size).

To align with previous work on intervention, I also looked at the effect of intervention quality in the standard condition and compared this with the yoked condition. Interestingly, the standard condition showed the exact opposite pattern—learners *choosing* good interventions gave less accurate hypotheses, and this difference increased reliably over time. The good learners were less accurate due to persistent belief in the prior (despite also learning the correct links), suggesting that the "good" interventions were designed partly to confirm, rather than challenge,

the prior belief. Results from the likelihood ratings support this interpretation. The standard condition, overall, maintained a stronger belief in the prior link than the yoked condition.

Finally, the patterns of competition from in Experiments 1 and 2 were replicated. Learners chose sets of causes to the exclusion of other sets that competed to explain the same covariation data, and the prior competed only with directly inconsistent cause. This latter result was absent from only the standard below-median condition, which also showed the weakest belief in the prior relation across learning, and thus, was least likely to be affected.

8.2 Implications for Causal Learning

The current experiments are most related to studies of causal structure learning, where people learn the causal relations between three or more events. Among the few studies on this topic, none consider the role of prior causal hypotheses that compete directly with the true causes. Thus, the results from Experiments 1-3 offered new insights into the mechanisms of concept restructuring, and in turn, added new constraints to current views of how people learn causal structure through observation and intervention.

8.2.1 Hypotheses and feedback

Given only evidence about the covariation of items, performance in a causal learning task is typically low (Lagnado & Sloman, 2004, 2006; Sobel & Kushnir, 2006; Steyvers et al., 2003; P. A. White, 2006). To improve learning, I reduced the set of possible causes from 32 to 5 and gave learners indirect feedback on their causal hypotheses. With these methods, participants showed considerable learning in all three experiments.

Steyvers et al. (2003) avoided feedback because their study involved the learning of many causal structures, one after another. If the correct answers were given at the end of each

structure, one could learn associative patterns between the data and correct choices without ever reasoning about cause and effect. Although Experiments 1-3 each involved only one structure, and participants were assured that the structure was NOT changing over time, feedback on specific causes may also lower the incentives to pay attention to the covariation trials. To avoid this, I told learners only how "good" their hypotheses were on an undefined scale ranging from Very weak, Weak, Good, to Very Good. They did not receive direct feedback on any particular link and were blind to the correspondence between points on the feedback scale and the number of correct links. The rationale for such feedback was that learners often make predictions based on their current hypotheses, which are confirmed or disconfirmed by their later observations. In other words, they usually take a more active learning approach between sets of observations.

Whether people treat feedback the same as confirming/disconfirming observations is an empirical question to be addressed in follow-up work. However, there is little reason to worry that learners were not reasoning causally about the observations, as Steyvers et al. were concerned. The no-change condition had no information except for the data bearing on which causes were more likely before feedback, yet their choices were biased toward the correct links. One-way ANOVAs predicting link choice (yes vs. no) with link type (A.B vs. A.C or C.B vs. D.A or D.B) as a factor showed a marginal effect of link for Experiment 1, $F(2, 46)=2.72$, $p=0.08$, and a significant effect of link for Experiment 2, $F(2, 54)=12.96$, $p<0.01$. Among the pairwise differences in link type, people were more likely to choose the correct links than the incorrect links (minus the prior) in both experiments. Hence, learners must have used a combination of the covariation data and the feedback to learn the true causal structure.

8.2.2 Incremental learning

Despite low performance in typical laboratory studies, people do have knowledge of complex, real-world causal systems (e.g., relations between diseases and symptoms). How do we learn these systems of causes? One idea is that the typical causal learning study, where participants have no prior knowledge for the target domain, is an inaccurate model of real world learning. More likely, people build their causal models one piece at a time, and many of their observations and interventions do not contribute *de novo* to causal learning, but instead are used to constrain and restructure elements from prior knowledge.

Fernbach and Sloman (2009) proposed a similar view, arguing that people learn sets of causes one (or few) at a time using simple heuristics. To support this claim, they showed that people have a systematic bias when learning causal chains, predicted by a learner accumulating causes in a piecemeal fashion. In their study, people observed a set of interventions chosen by the experimenter. In truth, property A caused B, and B caused C, each with a causal strength of 0.8. Hence, when A was turned "on," usually B and C occurred, which prompted learners to infer causes A.B and A.C. Additionally, when B was turned "on," C usually occurred, which led to a belief in B.C. Given these outcomes, a rational learner should step back and note that causes A.B and B.C are sufficient to explain the outcomes (A.C is not necessary to explain the relation between A and C), but most learners retained the full set of causes, consistent with an "accumulator" heuristic whereby people add single causes, as needed, to explain the current trial.

The heuristic view also explained how people revise their prior beliefs in response to conflicting evidence. For example, imagine a learner who has accumulated all three causes A.B, B.C, and A.C. But then, a trial in which A is turned "on" leads only to B. This trial calls into question the A.C relation, because if A causes C, then why was C not on as well? According to

the accumulator heuristic, when participants view this inconsistent trial *after* they have already inferred A.C, they will relinquish their belief in A.C. However, if this trial occurs *before* they infer A.C, and thus, is not yet inconsistent, it does not count against the A.C relation. Fernbach and Sloman (2009) confirmed this prediction. People had a stronger belief in A.C when the A "on" causing B only trial occurred before A.C was inferred, compared to after.

The heuristic view is consistent with the general claim of this thesis—that learning involves the accumulation of evidence, which sometimes calls into question our prior beliefs. However, it may not apply directly to Experiments 1-3 for two reasons. First, the learning trials in Experiments 1 and 2 were static observations, which lack the crucial time lapse between causes and potential effects, a major clue to causal structure (Lagnado & Sloman, 2004, 2006).

Second, in Experiment 3 observed causes were not *necessary* to lead to their effects (properties could also be "on" due to an unobserved background cause), but they were necessary in Fernbach and Sloman (2009). To see why this matters, note that in Experiment 3 turning A "on" may have co-occurred with C being "on," though in truth, A did not cause C. The accumulator model seems to predict that people will endorse the A.C relation as long as it is consistent with the most recent set of trials (i.e., is not contested by an A "on" and C "off" trial). Though technically possible, a more likely account of the few A.C hypotheses is that people tracked the low overall frequency of A co-occurring with C, and thus, tended to think this relation was untrue.

Another difference between the current work and the predictions of the heuristics view is that learners in Experiments 1-3 choose sets of causes to the exclusion of other sets (e.g., causes in the chain A.C+C.B, but not the common cause D.A+D.B). This seems to contradict the heuristic view, because for a learner merely accumulating causes, the belief in one cause (or set

of causes) should be independent of another. A heuristic learner could, in principle, show dependence between competing causes by shifting back and forth between the common cause and the causal chain, revising his beliefs in response to conflicting evidence, but this relies on a very specific pattern of trials. A more natural assumption is that sets of causes compete when they explain the same patterns of covariation (cf. Lu et al., 2008; Rehder & Milovanovic, 2007).

8.2.3 Constraints from prior knowledge

Prior knowledge offers many constraints that simplify the search for causal structure. For example, when told that murder rates correlate with ice cream consumption, our prior knowledge says there must be a common cause account rather than a direct relation between these arbitrary events. The constraints vary in abstractness—some are linked to a perceptual feature (e.g., aversive tastes are more likely to cause nausea than sounds; Garcia & Koelling, 1966), others depend on category knowledge (e.g., mental states do not usually cause physiological symptoms, Schultz et al., 2007), and still others are entirely abstract (e.g., a preference for simpler explanations, Lombrozo, 2007; or for necessary and sufficient causes, Lu et al., 2008; Rehder & Milovanovic, 2007).

The current findings suggest three new constraints: (1) Learners choose sets of causes to the exclusion of other sets that compete to explain the same covariation data; (2) Learners choose causes that compete with prior knowledge more often if they lead to different effects than the prior causes; (3) Learners relax their preference for the simplest causal structure when they learn new causes that compete with their prior beliefs. These constraints build naturally on current models, and I suggest possible ways to incorporate them below.

A recent model by Lu et al. (2008) predicts inter-link competition effects similar to those from Experiments 1-3 using two constraints on Cheng's (1997) power-PC theory. In short, the

model assumes that learners prefer one strong cause for a given effect, and disprefer many weak causes. Furthermore, if there are two possible causes, people assume that one is strong and the other is very weak. With these assumptions, Lu et al. improved the fit of power-PC to several data sets, confirming that people use a similar constraint when making causal inferences.

Rehder and Milovanovic (2007) also found evidence for competition. Their participants viewed sample data from a common cause structure and then gave likelihood estimates, which were fit to Causal Model Theory. Then, they saw the sample again and repeated the likelihood ratings. With the additional data, the ratings changed, and the direction of the change revealed the learners' biases. First, the estimated strength of background causes increased with more data, suggesting an initial preference for weak background causes. Second, the estimated strength of observed causes decreased with more data, suggesting an initial preference for strong observed causes. Together, these findings are consistent with Lu et al. (2008)—people assume that one cause (in this case, the observed cause) is relatively strong and all other potential causes (the background causes) are relatively weak.

How could we extend these models to account for competition among *sets* of causes? One solution is to modify the prior distributions over hypotheses, so that hypotheses with both competing explanations are always less likely than either explanation alone. This method is efficient but requires knowing which sets of causes form useful explanations, and then, which explanations compete. Though the modeler can easily identify these explanations and competition relations, a preferred method is to build this process into the model itself. For example, allow the model to build relevant explanations (from a set of atomic causal models), given the current observations or interventions, and then decide which explanations compete.

Hierarchical Bayesian models suggest a framework for this solution (e.g., Kemp, Goodman, & Tenenbaum, 2007; Kemp, Perfors, & Tenenbaum, 2007; Kemp & Tenenbaum, 2008; Tenenbaum, Griffiths, & Kemp, 2006; Tenenbaum, Griffiths, & Niyogi, 2007). In a hierarchical Bayesian model, each level of the hierarchy performs a type of inference. Consider the task of inferring the causal relations between a set of risk factors, symptoms, and diseases. Tenenbaum et al. (2006) suggested a model that uses observational data to make specific causal inferences at the lower level (e.g., that fatty diets cause lethargy), and simultaneously, uses the specific causal inferences to infer theoretical structures at a higher level. For example, a likely theory given trends in real-world medical data is that events in the category Risk Factors tend to cause events in the category Diseases, and that Diseases cause Symptoms. This type of theory is very useful if one wishes to predict, for example, whether a new disease is more likely to cause symptoms or risk factors (see Kemp, Tenenbaum, Niyogi, & Griffiths, 2009, for applications to other real-world domains).

Theoretical knowledge is not limited to categories and their relations; it can be anything that constrains the lower-level inferences. To explain the current data, one could build a theory that says which sets of causes form useful explanations and which explanations are in competition, based jointly on the correlational patterns in the data and principles of explanation generation (e.g., that correlations may be due to chains or common causes). A goal for this model would be to explain why causes compete less when they are not directly inconsistent with prior beliefs, as I found in Experiments 1-3. The details of such a model are not provided in this paper, but the hierarchical approach offers promise to understanding these constraints and others on causal structure learning.

8.3 Implications for Conceptual Change

This project brings together studies of causal inference and conceptual change by examining the role of specific prior beliefs when learning a new causal system. I took a modest first step, testing if and how learners would replace their prior belief in a direct causal relation with a slightly more complex structure (the true structure)—either a common cause or a causal chain. This technique shed light on the mechanisms of causal learning, which must also play a role in conceptual change. A major result was that learners assimilate the new causes with their prior knowledge and often do not revise their previous beliefs outright. This finding coheres with previous work on belief revision in scientific reasoning and theories of conceptual change.

8.3.1 Assimilating evidence and prior beliefs

Conceptual change occurs gradually over time as we make new observations and are exposed to ideas by our teachers and peers that conflict with prior beliefs. The process of interpreting such evidence and deciding what, if anything, to change about our current beliefs is a crucial part of cognitive development.

People respond to contradicting evidence in a number of ways, such as ignoring the data, holding the data in abeyance, reinterpreting the data, and sometimes (though least often) making changes in the theory the data contradicts (Chinn & Brewer, 1993). When theories are modified to accommodate the data, the changes are often peripheral, leaving the core assumption intact. The revisions in causal knowledge evident in Experiments 1-3 are closest to peripheral theory change, since people in the change condition opted against removing the prior concept outright, but showed weakened competition between the prior and some elements of the correct structure.

To what extent were the theory accommodation patterns in the current study reasonable, or scientifically valid? Previous developmental studies on the coordination of theory and data

offer conflicting viewpoints. Two representative sets of studies were conducted by D. Kuhn et al. (1988) and Koslowski (1996). The initial studies by D. Kuhn et al. addressed whether and to what extent people refer to covariation data when making causal inferences. The authors found consistent evidence that children (and adults, to a lesser extent) fail to use covariation evidence to support a causal inference, citing instead their knowledge of causal mechanisms. In addition, they found that when prior beliefs go against the covariation data, children tend to misinterpret or selectively ignore the data. In the rare cases when they do take the data into account, this happens only after they have revised their mechanistic theory to explain the data.

Koslowski (1996) took a different approach, arguing that the lack of attention to covariation data and focus on mechanisms were a plus, consistent with good scientific practice. She and her colleagues pointed out that small amounts of anomalous data typically do not prompt comprehensive theory revision, and especially not in science (see T. S. Kuhn, 1962). Instead, the initial theory is often considered to be a "working hypothesis," which can be revised down the road to account for new factors that were unspecified in the original theory. Accordingly, Koslowski's (1996) experiments focus on the types of evidence that lead to theory modification, as opposed to rejection. For example, when the instances that go against the theory have a common feature (e.g., advertising helps to sell textbooks, except for books with expensive color images), as opposed to multiple unrelated features (e.g., … except for books with color images or quickly outdated research), then the theory is more likely to be modified than rejected.

The current studies are interesting in light of this debate. Despite the finding that most people revised their theory of the ecosystem, the changes that took place are not easily cast as the elaboration of a "working hypothesis," as Koslowski describes. Consider Experiment 1, for example, where the direct cause and common cause explanations have no links in common, and

thus, are completely at odds. To revise the prior belief, one cannot simply make an exception category where for some cases, A does not cause B, because A *never* causes B. Furthermore, the data unambiguously suggest a common cause, with equal evidence for relations D.A and D.B. Yet, learners in the change condition assimilated the prior belief and the true structure by choosing the less inconsistent relation D.A more when choosing the prior relation A.B. This finding was replicated in both Experiments 2 and 3.

Based on these finding, it seems that people not only ignore and misinterpret theory-inconsistent data (as found by D. Kuhn et al., 1988), they actually form hypotheses that blend the theory and evidence, as if to bring them into agreement. Furthermore, because the direct cause and the common cause are directly incompatible, this blending is not in the service of editing a theory to account for exceptions; assimilation occurs even in cases where background knowledge and covariation data provide no reasonable justification for doing so.

8.3.2 Views of conceptual change

To understand the scope of the current work, it's important to ask what types of conceptual change, if any, are being captured by the ecosystem structures from Experiments 1-3. Participants in the change condition learned about just one new property in the target learning phase, and were prompted to re-explain just one correlation. According to some views, these steps alone do not qualify as conceptual change, even if they may play a role in the conceptual change process. In this section, I discuss two prominent views on what is conceptual change, in the proper sense, to put the current results into context.

Carey (1988, 1999) has been careful to separate conceptual change from other varieties of knowledge restructuring. Earlier I made reference to Carey's distinction between weak and strong restructuring (where strong refer to conceptual change in the history of science sense),

arguing how both may be realized by changes in causal networks and that doing so blurs their distinction. Nonetheless, it remains true that in order to qualify as strong restructuring, a given incident of *causal restructuring* must result in (1) changes at the level of individual concepts, (2) changes in the nature of explanation, and (3) changes in the domain to which those explanation apply. Evaluating these criteria is relatively simple. The first is at least partly satisfied, because participants learned new concepts and new relations that redefined the ecosystem's internal structure. The latter two are probably not satisfied, since the "nature of explanation" is undefined without multiple explanations to generalize across, and the one true explanation was only ever intended to account for properties A-D. Hence, in Carey's view, the current studies represent only a fraction of the conceptual development necessary to count as genuine conceptual change.

A different view is given by (Vosniadou, 1994), who argues for a distinction between types of conceptual change that vary in difficulty, but no firm exclusion of the weaker types. First, a distinction is made between *framework theories*, which contain the core assumptions about the nature of events in a domain (e.g., for physics, the property "heat" is transferrable), and *specific theories*, which concern the specific interactions between events in the domain (e.g., heat can be transferred across objects by direct contact). Second, a distinction is made between knowledge *enrichment*—the addition of new information to extant conceptual structures (akin to Carey's weak restructuring), and knowledge *revision*—the restructuring of one's current theoretical knowledge. Within this pair of distinctions, Vosniadou says that the revision of framework theories is more difficult than revision of specific theories, though both are types of conceptual change.

According to Vosniadou, it would seem that the changes in Experiments 1-3 were in fact conceptual change, though perhaps of a weaker form. Participants did restructure their view of

the ecosystem to some extent, even if they did not engage in the more difficult type of restructuring, which presumably would imply revision in one's core assumptions about ecology; e.g., that underwater animals and plants are co-dependent.

Ultimately, it may be difficult to capture strong restructuring (according to either theorist) in the course of a standard one-hour laboratory experiment (though see Chinn & Brewer, 1993, p.12, for a small list of exceptions). Future work should attempt to operationalize the stronger versions conceptual change within a causal model view (see Lucas & Griffiths, 2010, for work in this direction), and then search for task manipulations that either promote or discourage it.

8.4 Implications for Classroom Learning—Discovery vs. Guided Instruction

A benefit to studying the cognitive mechanisms of causal learning and revision is that new ideas may result for how to foster conceptual change in students. This benefit was apparent in Experiment 3, where I found that learners intervening on properties chosen by a teacher (the computer) showed more concept revision than learners choosing their own interventions. Though preliminary, this finding has implications for the use of active learning tasks in the classroom.

An important job for educators is to provide a balance between guided instruction and self-discovery. Many previous studies have compared these learning styles directly to see if one leads to better learning and transfer, in the general case. In "minimally guided" learning (also called "discovery" learning), the teacher poses a problem but provides few suggestions for how to obtain the solution. The student must construct their knowledge of the domain from the bottom up, including new rules, concepts, and solution procedures. In guided instruction, the exact opposite is true. Teachers explain all of the necessary rules and concepts upfront, and often give example problems with completed solutions.

In a general way, the standard and yoked intervention conditions from Experiment 3 map onto discovery and guided instruction, respectively. A study by Vollmeyer, Burns, and Holyoak (1996) makes this analogy concrete. In their task, participants could manipulate aspects of the water quality in a fish tank to bring about changes in the fish populations (strangely, I was not aware of this study before designing Experiments 1-3). In one experiment, people learned the relations between the water quality and the fish population by either (a) attempting to bring about a specific number of each type of fish, or (b) freely exploring the relations between the water quality and the fish. Results showed benefits of the latter condition (the minimal guidance condition). Learners freely exploring the relations between water quality and fish showed more knowledge of the correct relations and were also more accurate in a transfer task where both conditions had to produce a new set of fish populations.

Note the similarities of the Vollmeyer et al. study to the current work. The free exploration condition is nearly identical to the standard intervention condition of Experiment 3, and the yoked condition is similar to the alternative condition in that the experimenter (not the learner) chose the task on each trial. Interestingly, however, this analogy suggests the opposite pattern of the results in Experiment 3, i.e., that learners choosing their own interventions would do better. In fact, this condition was less likely to choose the correct links, and more likely to retain their belief in the incorrect prior. What accounts for these opposing findings?

One possibility is that direct instruction helps more when the learning task involves the revision of prior knowledge (as it did in Experiment 3 but not in Vollmeyer et al). Consistent with this idea, an earlier study by B. Y. White (1984) found that students with a specific computer task chosen by the experimenter benefited more when learning about Newtonian dynamics, a domain where students are infamously hindered by their intuitive prior knowledge

(e.g., Smith III et al., 1994). Much like the interventions on property A in Expeirment 3, the specific task in White's study was designed to draw learners' attention to their faulty prior beliefs so that they see the flaws and attempt revision.

Further evidence for the role of prior knowledge comes from the causal induction literature. The typical finding in studies where learners *do not* have prior knowledge of the causal system (all previous studies) is that the standard intervention condition outperforms the yoked condition (Lagnado & Sloman, 2004; Sobel & Kushnir, 2006). One could test the role of prior beliefs by redoing Experiment 3 with both change and no-change conditions, predicting an interaction between prior knowledge (yes/no) and intervention condition.

A different possibility is that basic cognitive demands (which may differ across experiments) modulate the effects of guidance and discovery (cf. Kirschner, Sweller, & Clark, 2006). Discovery tasks are in essence a "dual-task," with both learning and search processes competing for working memory. When search is intensive, learning suffers. This may explain why in Experiment 3 the yoked condition outperformed the standard condition, if removing the need to choose (i.e., search for useful interventions) frees up learning resources that may be used to analyze and interpret the intervention outcomes.

Though attractive, this interpretation seems at odds with the findings of Sobel and Kushnir (2006) and Lagnado and Sloman (2004), who showed a benefit to choosing interventions. Yet again, participants in those studies did not evaluate their chosen interventions in light of conflicting prior beliefs. It may be that in the absence of prior beliefs, working memory load is sufficiently low so that the benefits of choosing one's own interventions dominate (e.g., knowing the organizing structure behind the interventions; see Sobel & Kushnir,

2006). This conclusion awaits further data clarifying the distinct roles of intervention search and organization strategies.

Regardless of how one interprets the findings of Experiment 3, it is nonetheless important to understand how different task factors (such as direct vs. indirect instruction) may promote or discourage conceptual change. Future research should aid to a better understanding of these issues, and more generally, relate causal intervention to other interactive tasks from education research, such as computer simulation and experimentation, which have been shown to promote abstract generalizations (e.g., Clement, 1993; Goldstone & Son, 2005) and improve classroom learning (e.g., Roschelle, Pea, Hoadley, Gordin, & Means, 2000; Thornton & Sokoloff, 1990).

CHAPTER 9: CONCLUSIONS

This thesis brings together research on causal learning and conceptual change. Though the empirical methods and analytic tools used to understand these topics differ dramatically, the underlying theoretical constructs have much in common. Both areas emphasize the role of causal/theoretical knowledge (and revisions thereof) in concept development, suggesting a path to unification.

Toward this aim, Experiments 1 and 2 presented a new paradigm for the study of causal learning *and restructuring*. Learners' prior beliefs about the causal relations in a domain affected their hypotheses as they begin to infer the correct causes, in both helpful and hurtful ways. First, when the prior learning suggested evidence against some of the incorrect causes, this helped learners to focus on the correct relations during the primary learning phase. Second, the prior causal beliefs were difficult to give up, and as a consequence, they biased learners away from the correct causes that competed to explain the same effects.

Experiment 3 showed that intervening (or experimenting) on the causal domain affected the concept restructuring process in different ways, depending on what interventions were chosen and by whom. People choosing their own interventions revealed a confirmation bias to preserve their prior beliefs, but those implementing the same interventions used the more diagnostic intervention sets to disconfirm their prior beliefs. Taken together, these studies represent the beginnings of a larger research effort to use the analytic tools from causal induction to reveal the mechanisms behind larger shifts in knowledge, as evidenced by developing children and experts.

REFERENCES

Ahn, W., Kim, N. S., Lassaline, M. E., & Dennis, M. J. (2000). Causal Status as a Determinant of Feature Centrality. *Cognitive Psychology*, *41*(4), 361–416.

Atran, S. (1995). Causal constraints on categories and categorical constraints on biological reasoning across cultures. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal cognition: A multidisciplinary debate* (pp. 205–233). Oxford, England: Clarendon Press.

Baillargeon, R., & Goswami, U. (2002). The acquisition of physical knowledge in infancy: A summary in eight lessons. In *Blackwell handbook of childhood cognitive development* (pp. 46-83). Oxford, England: Blackwell.

Blessing, S. B., & Ross, B. H. (1996). Content effects in problem categorization and problem solving. *Learning, Memory*, *22*(3), 792–810.

Buehner, M. J., Cheng, P. W., & Clifford, D. (2003). From covariation to causation: A test of the assumption of causal power. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(6), 1119–1140.

Carey, S. (1988). Reorganization of knowledge in the course of acquisition. In S. Strauss (Ed.), *Ontogeny, Phylogeny, and Historical Development* (pp. 1–27). Norwood, NJ: Ablex.

Carey, S. (1999). Sources of conceptual change. In E. K. Scholnick, K. Nelson, S. A. Gelman, & P. Miller (Eds.), *Conceptual development: Piaget's legacy* (pp. 293-326). Hillsdale, NJ: Erlbaum.

Carey, S. (1985). *Conceptual Change in Childhood.* Cambridge, MA: Bradford Books, MIT Press.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, *4*(1), 55–81.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*(2), 367–405.

Chi, M. T., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive science*, *5*(2), 121–152.

Chi, M. T., Slotta, J. D., & De Leeuw, N. (1994). From things to processes: A theory of conceptual change for learning science concepts. *Learning and Instruction*, *4*(1), 27–43.

Chinn, C. A., & Brewer, W. F. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. *Review of Educational Research*, *63*(1), 1.

Clement, J. (1993). Using bridging analogies and anchoring intuitions to deal with students' preconceptions in physics. *Journal of Research in Science Teaching*, *30*(10), 1241–1257.

Cohen, L. B., & Oakes, L. M. (1993). How infants perceive a simple causal event. *Developmental Psychology*, *29*, 421–421.

disessa, A. A., Gillespie, N. M., & Esterly, J. B. (2004). Coherence versus fragmentation in the development of the concept of force. *Cognitive Science*, *28*(6), 843–900.

diSessa, A. A. (1993). Toward an epistemology of physics. *Ethics & Behavior*, *10*(2), 105–225.

diSessa, A. A., & Sherin, B. L. (1998). What changes in conceptual change? *International Journal of Science Education*, *20*(10), 1155–1191.

Doumas, L. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review*, *115*(1), 1–43.

Elman, J. L., Bates, E., Johnson, M., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: Connectionism in a developmental framework*. Cambridge, MA: MIT Press.

Fernbach, P. M., & Sloman, S. A. (2009). Causal Learning With Local Computations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(3), 16.

Flavell, J. H. (1996). Piaget's legacy. *Psychological Science*, *7*(4), 200–203.

Fugelsang, J. A., Stein, C. B., Green, A. E., & Dunbar, K. N. (2004). Theory and Data Interactions of the Scientific Mind: Evidence From the Molecular and the Cognitive Laboratory. *Canadian Journal of Experimental Psychology*, *58*(2), 86–95.

Garcia, J., & Koelling, R. A. (1966). Relation of cue to consequence in avoidance learning. *Psychonomic Science. Vol 4(3)*, 123-124.

Goldstone, R. L. (1996). Isolated and interrelated concepts. *Memory & Cognition*, *24*, 608–628.

Goldstone, R. L., & Rogosky, B. J. (2002). Using relations within conceptual systems to translate across conceptual systems. *Cognition*, *84*(3), 295–320.

Goldstone, R. L., & Son, J. Y. (2005). The transfer of scientific principles using concrete and idealized simulations. *Journal of the Learning Sciences*, *14*(1), 69–110.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*(1), 108–154.

Gopnik, A. (1996). The post-Piaget era. *Psychological Science*, *7*(4), 221–225.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A Theory of Causal Learning in Children: Causal Maps and Bayes Nets. *Psychological Review*, *111*(1), 3–32.

Gopnik, A., Sobel, D. M., Schulz, L. E., & Glymour, C. (2001). Causal learning mechanisms in very young children: Two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental Psychology*, *37*(5), 620–629.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*(4), 334–384.

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review*, *116*(4), 661–716.

Gwiazda, J., Ong, E., Held, R., & Thorn, F. (2000). Myopia and ambient night-time lighting. *Nature*, *404*, 144.

Hume, D. (1978). *A treatise of human nature*. Oxford, England: Oxford University Press.

Inagaki, K., & Hatano, G. (1993). Young children's understanding of the mind-body distinction. *Child Development*, *64*(5), 1534–1549.

Ioannides, C., & Vosniadou, S. (2002). The changing meanings of force. *Cognitive Science Quarterly*, *2*(1), 5–62.

Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs: General and Applied*, *79*((1, Whole No. 594)).

Kao, S. F., & Wasserman, E. A. (1993). Assessment of an information integration account of contingency judgment with examination of subjective cell importance and method of information presentation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(6), 1363–1386.

Kaplan, A. S., & Murphy, G. L. (2000). Category learning with minimal prior knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(4), 829–846.

Keil, F. C. (1992). The origins of an autonomous biology. In M. R. Gunnar & M. Maratsos (Eds.), *Modularity and constraints in language and cognition* (Vol. 25, pp. 103–138). Hillsdale, NJ: Erlbaum.

Kelemen, D. (1999). Beliefs about purpose: On the origins of teleological thought. In M. Corballis & S. Lea (Eds.), *The descent of mind: Psychological perspectives on hominid evolution* (pp. 278–294). Oxford, England: Oxford University Press.

Kemp, C., Goodman, N. D., & Tenenbaum, J. B. (2007). Learning causal schemata. In *Proceedings of the 29th Annual Cognitive Science Society.* (pp. 389–394).

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, *10*(3), 307–321.

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*, *105*(31), 10687.

Kemp, C., Tenenbaum, J. B., Niyogi, S., & Griffiths, T. L. (2009). A probabilistic model of theory formation. *Cognition*.

Kim, N. S., & Ahn, W. (2002a). Clinical psychologists' theory-based representations of mental disorders predict their diagnostic reasoning and memory. *Journal of Experimental Psychology: General*, *131*(4), 451–476.

Kim, N. S., & Ahn, W. K. (2002b). The influence of naive causal theories on lay concepts of mental illness. *The American journal of psychology*, *115*(1), 33–65.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational psychologist*, *41*(2), 75–86.

Koslowski, B. (1996). *Theory and evidence: The development of scientific reasoning*. The MIT Press.

Kruschke, J. K. (1999). ALCOVE: An exemplar-based connectionist model ot category learning. *Psychological Review*, *99*(1), 22-44.

Kuhn, D., Amsel, E., O'Loughlin, M., & Beilin, H. (1988). *The Development of Scientific Thinking Skills*. Academic Press.

Kuhn, T. S. (1970). *The structure of scientific revolutions, International Encyclopedia of Unified Science, vol. 2, no. 2*. Chicago: University of Chicago Press.

Lagnado, D. A., & Sloman, S. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(4), 856–876.

Lagnado, D. A., & Sloman, S. A. (2006). Time as a guide to cause. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(3), 451–460.

Leslie, A. M., Gallistel, C. R., & Gelman, R. (2008). Where integers come from. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *Where Integers Come From* (Vol. 3, pp. 109–138). Oxford, England: Oxford University Press.

Leslie, A. M., & Keeble, S. (1987). Do six-month-old infants perceive causality? *Cognition*, *25*(3), 265–288.

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology*, *55*(3), 232–257.

Lombrozo, T., & Carey, S. (2006). Functional explanation and the function of explanation. *Cognition*, *99*(2), 167–204.

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review*, *115*(4), 955–984.

Lucas, C. G., & Griffiths, T. L. (2010). Learning the Form of Causal Relationships Using Hierarchical Bayesian Models. *Cognitive Science*, *34*(1), 113–147.

Luhmann, C. C., & Ahn, W. (2007). BUCKLE: A model of unobserved cause learning. *Psychological Review*, *114*(3), 657–676.

Marsh, J. K., & Ahn, W. K. (2006). Order effects in contingency learning: The role of task complexity. *Memory & Cognition*, *34*(3), 568.

Mayer, R. E. (1981). Frequency norms and structural analysis of algebra story problems into families, categories, and templates. *Instructional Science*, *10*(2), 135–175.

McCloskey, M. (1983). Naive theories of motion. In D. Gentner & A. L. Stevens (Eds.), *Mental models* (pp. 299–324). Hillsdale, NJ: Erlbaum.

Meder, B., Hagmayer, Y., & Waldmann, M. R. (2009). The role of learning data in causal reasoning about observations and interventions. *Memory & Cognition*, *37*(3), 249.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review*, *100*(2), 254–278.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207–238.

Medin, D. L., & Shoben, E. J. (1988). Context and structure in conceptual combination. *Cognitive Psychology*, *20*(2), 158–190.

Michotte, A. (1963). *The Perception of Causality* (First American Edition.). Basic Books.

Minda, J. P., & Smith, J. D. (2002). Comparing Prototype-Based and Exemplar-Based Accounts of Category Learning and Attentional Allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(2), 275–292.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*(3), 289–316.

Murphy, G. L. (2002). *The Big Book of Concepts*. The MIT Press.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57.

Pearl, J. (2000). *Causality: models, reasoning, and inference*. Cambridge, England: Cambridge University Press.

Piaget, J. (1952). *The Origins of Intelligence in Children* (1st ed.). International Universities Press.

Preston, J., & Epley, N. (2009). Science and God: An automatic opposition between ultimate explanations. *Journal of Experimental Social Psychology*, *45*(1), 238–241.

Quinn, G. E., Shin, C. H., Maguire, M. G., & Stone, R. A. (1999). Myopia and ambient lighting at night. *Nature*, *399*(6732), 113-114.

Rehder, B. (2003a). A causal-model theory of conceptual representation and categorization. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *29*(6), 1141–1159.

Rehder, B. (2003b). Categorization as causal reasoning. *Cognitive Science*, *27*(5), 709–748.

Rehder, B. (2006). When similarity and causality compete in category-based property induction. *Memory & Cognition*, *34*, 3–16.

Rehder, B., & Kim, S. W. (2006). How causal knowledge affects classification: A generative theory of categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(4), 659.

Rehder, B., & Milovanovic, G. (2007). Bias toward sufficiency and completeness in causal explanations. In *Proceedings of the 29th Annual Conference of the Cognitive Science Society* (p. 1843).

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573–605.

Roschelle, J. M., Pea, R. D., Hoadley, C. M., Gordin, D. N., & Means, B. M. (2000). Changing how and what children learn in school with computer-based technologies. *Children and Computer Technology*, *10*(2), 76–101.

Ross, B. H. (1984). Remindings and their effects in learning a cognitive skill. *Cognitive Psychology*, *16*(3), 371–416.

Ross, B. H., & Kennedy, P. T. (1990). Generalizing from the use of earlier examples in problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(1), 42–55.

Rottman, B. M., Ahn, W. K., & Luhmann, C. C. (in press) When and how do people reason about unobserved causes? In P. Illari, F. Russo, & J. Williamson (Eds.), *Causality in the Sciences*. New York: Oxford University Press.

Scheines, R., Spirtes, P., Glymour, C. N., & Meek, C. (1994). *TETRAD II: Tools for causal modeling*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Schulz, L. E., Bonawitz, E. B., & Griffiths, T. L. (2007). Can Being Scared Cause Tummy Aches? Naive Theories, Ambiguous Evidence, and Preschoolers' Causal Inferences. *Developmental Psychology*, *43*(5), 1124–1139.

Sloman, S. A. (1994). When explanations compete: the role of explanatory coherence on judgments of likelihood. *Cognition*, *52*(1), 1–21.

Sloman, S. A., Love, B. C., & Ahn, W. K. (1998). Feature centrality and conceptual coherence. *Cognitive Science*, *22*(2), 189–228.

Smith III, J. P., diSessa, A. A., & Roschelle, J. (1994). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *Journal of the Learning Sciences*, *3*(2), 115–163.

Sobel, D. M., & Kushnir, T. (2006). The importance of decision making in causal learning from interventions. *Memory & Cognition*, *34*(2), 411.

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental Science*, *10*(1), 89–96.

Spirtes, P., Glymour, C., & Schienes, R. (1993). *Causation, Prediction, and Search.* New York: Springer-Verlag.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E. J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, *27*(3), 453–489.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*(7), 309–318.

Tenenbaum, J. B., Griffiths, T. L., & Niyogi, S. (2007). Intuitive theories as grammars for causal inference. In A. Gopnik & L. E. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 301–322). Oxford University Press.

Thaden-Koch, T. C., Dufresne, R. J., & Mestre, J. P. (2006). Coordination of knowledge in judging animated motion. *Physical Review Special Topics-Physics Education Research*, *2*(2), 20107.

Thornton, R. K., & Sokoloff, D. R. (1990). Learning motion concepts using real-time microcomputer-based laboratory tools. *American Journal of Physics*, *58*(9), 858–867.

Vollmeyer, R., Burns, B., & Holyoak, K. J. (1996). The impact of goal specificity on strategy use and the acquisition of problem structure. *Cognitive Science*, *20*(1), 75–100.

Vosniadou, S. (1994). Capturing and modeling the process of conceptual change. *Learning and Instruction*, *4*(1), 45–69.

Vosniadou, S. (2008). *International handbook of research on conceptual change*. New York: Routledge.

Vosniadou, S., & Brewer, W. F. (1992). Mental models of the earth: A study of conceptual change in childhood. *Cognitive Psychology*, *24*(4), 535–585.

Waldmann, M. R., & Hagmayer, Y. (2006). Categories and causality: The neglected direction. *Cognitive Psychology*, *53*(1), 27–58.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *The Quarterly Journal of Experimental Psychology*, *12*(3), 129–140.

Wellman, H. M., & Gelman, S. A. (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology*, *43*(1), 337–375.

White, B. Y. (1984). Designing Computer Games to Help Physics Students Understand Newton's Laws of Motion. *Cognition and Instruction*, *1*(1), 69–108.

White, P. A. (2006). How well is causal structure inferred from cooccurrence information? *European Journal of Cognitive Psychology*, *18*(3), 454.

Woodward, A. L. (2005). The infant origins of intentional understanding. In R. V. Kail (Ed.), *Advances in Child Development and Behavior* (Vol. 33, pp. 229–262). Oxford, England: Elsevier.

APPENDIX

According to Bayes' rule, $P(g \mid y,a) = \frac{P(y \mid g,a)P_s(g \mid a)}{P(y \mid a)}$. However, a can be dropped from $P_s(g|a)$, because this represents the prior belief in graphs g, before the outcome of intervention a has been observed. Hence, $P(g|y,a)$ can be substituted in Equation 2 as follows,

$$\langle H(a) \rangle = -\sum_y P(y \mid a) \sum_g \frac{P(y \mid g,a)P_s(g)}{P(y \mid a)} \log \frac{P(y \mid g,a)P_s(g)}{P(y \mid a)}, \tag{4}$$

which reduces to $$\langle H(a) \rangle = -\sum_y \sum_g P(y \mid g,a)P_s(g) \log \frac{P(y \mid g,a)P_s(g)}{P(y \mid a)}. \tag{5}$$

Finally, by Equation 3, $$\langle H(a) \rangle = -\sum_y \sum_g P(y \mid g,a)P_s(g) \log \frac{P(y \mid g,a)P_s(g)}{\sum_g P(y \mid g,a)P_s(g)}. \tag{6}$$

Computing uncertainty requires only the terms $P(y|g,a)$ and $P_s(g)$.